



Introduction à la FOUILLE DE TEXTES et positionnement de l'offre logicielle

Patrice Bellot (Aix-Marseille Univ, CNRS)

patrice.bellot@cnrs-dir.fr

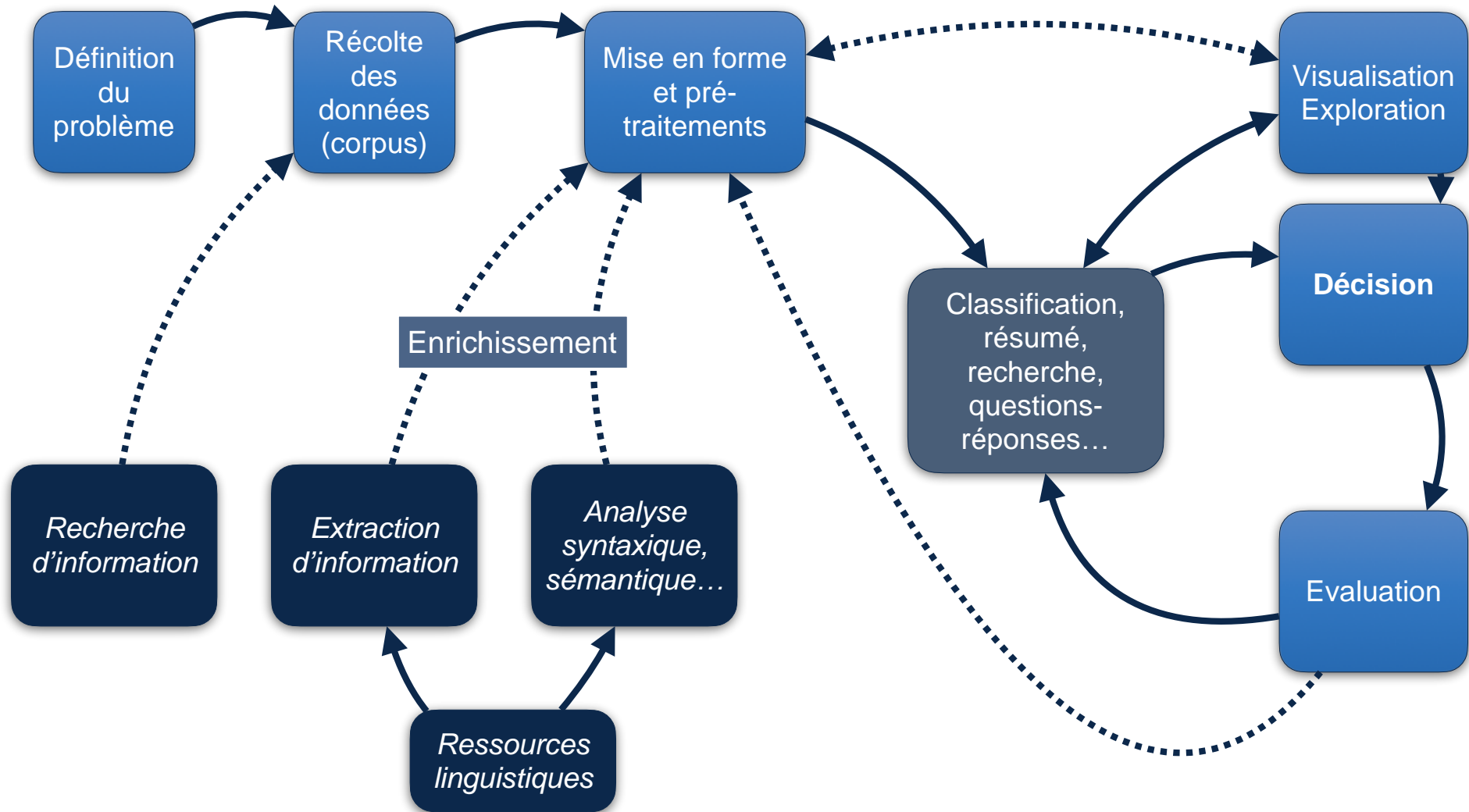
ANF TDM — novembre 2021

QU'EST-CE QUE LA FOUILLE DE TEXTES ?

Le croisement de plusieurs domaines

- L'analyse et la fouille de données (Data Mining)
- Le traitement automatique des langues
- La recherche et l'extraction d'information

Un processus de fouille de textes



QUELQUES CARACTÉRISTIQUES...

Quelle que soit la nature des données :

- Structures peu normalisées, formats variés
- Les V du Big Data : Volume, véracité, variabilité, valeur, vitesse

Documents, textes et langues :

- Données hétérogènes et multimodales
- Multilinguisme (variations : lexicque et terminologie, syntaxe...)
- La langue est ambiguë : polysémie, contexte, interprétation...

Des difficultés
génériques.

Des standards
nécessaires.

Des solutions
à partager.

Des solutions, des verrous : ingénierie et recherche

Des données, des
méta-données et des
formats : éléments
pour une mise en
forme

L'encodage des documents

Virtual Reality (2006) 10:135-147
DOI 10.1007/s10055-006-0048-0

ORIGINAL ARTICLE

The design and realization of CoViD: a system for collaborative virtual 3D design

Wolfgang Stuerzlinger · Loufouz Zaman ·
Andriy Pavlovych · Ji-Young Oh

Received: 14 March 2006 / Accepted: 12 May 2006 / Published online: 19 September 2006
© Springer-Verlag London Limited 2006

Abstract Many important decisions in the design process are made during fairly early on, after designers have presented initial concepts. In many domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today's systems and tools do not support this, as computer systems typically support only a single user and computer-aided design tools require significant training. This paper presents the design of a new system to facilitate a collaborative 3D design process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D design systems. Then, we introduce the new system, which consists of two main parts. The first part is an easy-to-use conceptual 3D design tool that can be used productively even by naive users. The tool provides novel interaction techniques that support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rapid creation and modification of 3D models. The second part is a novel infrastructure for collaborative work, which offers an interactive table and several large interactive displays in a semi-immersive setup. It is designed to support multiple users working together. This infrastructure also includes novel pointing devices that work both as a stylus and a remote pointing device. The combination of the (modified) design tool with the collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present an evaluation of the system against the guidelines for collaborative 3D design. Finally, we present results of a preliminary user study, which asked naive users to collaborate in a 3D design task on the new system.

Keywords Collaborative design · 3D design · Collaborative virtual reality

1 Introduction

Today, digital 3D models are critical in many domains, such as architecture and urban planning, all kinds of industrial design, the entertainment industry, and many engineering applications. Many of the important decisions surrounding a design are made in the initial phases, after the designer(s) have proposed a first version of the design. There, typically in a meeting, the stakeholders in the project get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the proposed designs to explore the design space better. Today's design tools and computer infrastructure do not support such activities well, as computer systems typically support only a single user and computer-aided design tools require significant training. Traditional tools for 3D design require a large amount of training. Part of this is based on the fact that

W. Stuerzlinger (✉) · L. Zaman · A. Pavlovych
York University, Toronto, Canada
URL: <http://www.cs.yorku.ca/~wolfgang>
URL: <http://www.cs.yorku.ca/~zaman>
URL: <http://www.cs.yorku.ca/~andryp>

J.-Y. Oh
University of Arizona, Tucson, AZ, USA
e-mail: jyoh@optics.arizona.edu

Springer

```
(base) [ Patrice Mac-Pro-de-Patrice ~ ] more /Users/Patrice/Downloads/ark_67375_
"/Users/Patrice/Downloads/ark_67375_VQC-MZJ5GK10-Q/PDF/10055_2006_Article_48.pdf
%PDF-1.3
<<
  /Type /Page
  /Contents 10 0 R
  /Parent 1 0 R
  /Resources <<
    /Font <<
      /F1 <<
        /Type /Font
        /Subtype /Type1
        /BaseFont /Helvetica
      >>
    >>
  >>
  /MediaBox [ 0 0 595 842 ]
  /CropBox [ 0 0 595 842 ]
  /Rotate 0
  /PageLabels <>
endobj
xref
1 2
0000000000 1 n
0000000009 1 n
trailer
<<
  /Size 2
  /Root 1 0 R
  /Info <>
  /Pages 1 0 R
endobj
startxref
9
%%EOF
```

PDF to text

Evitons les PDF...



L'encodage des caractères

<ul style="list-style-type: none"> ✓ Par défaut Occidental (ISO Latin 1) Occidental (Mac OS Roman)
Unicode (UTF-8)
<ul style="list-style-type: none"> Japonais (Shift JIS) Japonais (ISO 2022-JP) Japonais (EUC) Japonais (Shift JIS X0213)
<ul style="list-style-type: none"> Chinois traditionnel (Big 5) Chinois traditionnel (Big 5 HKSCS) Chinois traditionnel (Windows, DOS)
<ul style="list-style-type: none"> Coréen (ISO 2022-KR) Coréen (Mac OS) Coréen (Windows, DOS)
<ul style="list-style-type: none"> Arabe (ISO 8859-6) Arabe (Windows)
<ul style="list-style-type: none"> Hébreu (ISO 8859-8) Hébreu (Windows)
<ul style="list-style-type: none"> Grec (ISO 8859-7) Grec (Windows)
<ul style="list-style-type: none"> Cyrillique (ISO 8859-5) Cyrillique (Mac OS) Cyrillique (KOI8-R) Cyrillique (Windows) Ukrainien (KOI8-U)
Thaïlandais (Windows, DOS)
<ul style="list-style-type: none"> Chinois simplifié (GB 2312) Chinois simplifié (HZ GB 2312) Chinois (GB 18030)
<ul style="list-style-type: none"> Europe centrale (ISO Latin 2) Europe centrale (Mac OS) Europe centrale (Windows Latin 2)
<ul style="list-style-type: none"> Vietnamien (Windows)
<ul style="list-style-type: none"> Turc (ISO Latin 5) Turc (Windows Latin 5)
<ul style="list-style-type: none"> Europe centrale (ISO Latin 4) Balte (Windows)

Dec	Hex	Char	Dec	Hex	Char
64	40	@	96	60	'
65	41	A	97	61	a
66	42	B	98	62	b
67	43	C	99	63	c
68	44	D	100	64	d
69	45	E	101	65	e
70	46	F	102	66	f
71	47	G	103	67	g
72	48	H	104	68	h
73	49	I	105	69	i
74	4A	J	106	6A	j
75	4B	K	107	6B	k
76	4C	L	108	6C	l
77	4D	M	109	6D	m
78	4E	N	110	6E	n
79	4F	O	111	6F	o
80	50	P	112	70	p
81	51	Q	113	71	q

ASCII (années 1960, sur 7 bits)
American Standard Code for
Information Interchange

Dec	Hex	Char	Dec	Hex	Char
7	00A8	À	7	00AC	Ä
7	00B8	à	7	00BC	ä
7	00C8	È	7	00CC	È
7	00D8	è	7	00DC	è
7	00E8	È	7	00EC	è
7	00F8	ø	7	00FC	ÿ

ISO-Latin 1 sur 8 bits
(ISO/CEI 8859)

<https://unicode.org/emoji/charts/full-emoji-list.html>
<https://home.unicode.org>

face-smiling														
N°	Code	Browser	Appl	Goog	FB									
940	U+1F600	á	é	ή	ί	ύ	α	β	γ	δ	ε	☺	☺	☺
950		ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο			
960		π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω			
970		ϊ	ϋ	ό	ύ	ώ	κ	ε	θ	Υ	Υ	☺	☺	☺
980		ÿ	φ	ω	χ	Ϟ	ϟ	Ϡ	ϡ	Ϣ	ϣ			
990		Ϝ	ϝ	Ϟ	ϟ	Ϡ	ϡ	Ϣ	ϣ	Ϥ	ϥ	☺	☺	☺
1000		Ϧ	ϧ	Ϩ	ϩ	Ϫ	ϫ	Ϭ	ϭ	Ϯ	ϯ			
1010		ϰ	ϱ	ϲ	ϳ	ϴ	ϵ	϶	Ϸ	ϸ	Ϲ			
1020		ϻ	ϼ	Ͻ	Ͼ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	☺	☺	☺
1030		Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ	☺	☺	☺
1040		Α	Β	Γ	Δ	Ε	Ζ	Ζ	Η	Θ	Ι			
1050		Κ	Λ	Μ	Ν	Ο	Π	Ρ	Σ	Τ	Υ			
1060		Φ	Χ	Ψ	Ω	Ω	Ω	Ω	Ω	Ω	Ω			
1070		Ю	Я	а	б	в	г	д	е	ж	з			
1080		и	й	к	л	м	н	о	п	р	с			
1090		т	у	ф	х	ц	ч	ш	щ	ъ	ы			
1100		ь	э	ю	я	è	ë	ĥ	í	é	s			
1110		i	ı	j	љ	њ	ќ	ћ	ѝ	џ	џ			
1120		ω	w	џ	џ	џ	џ	џ	џ	џ	џ			

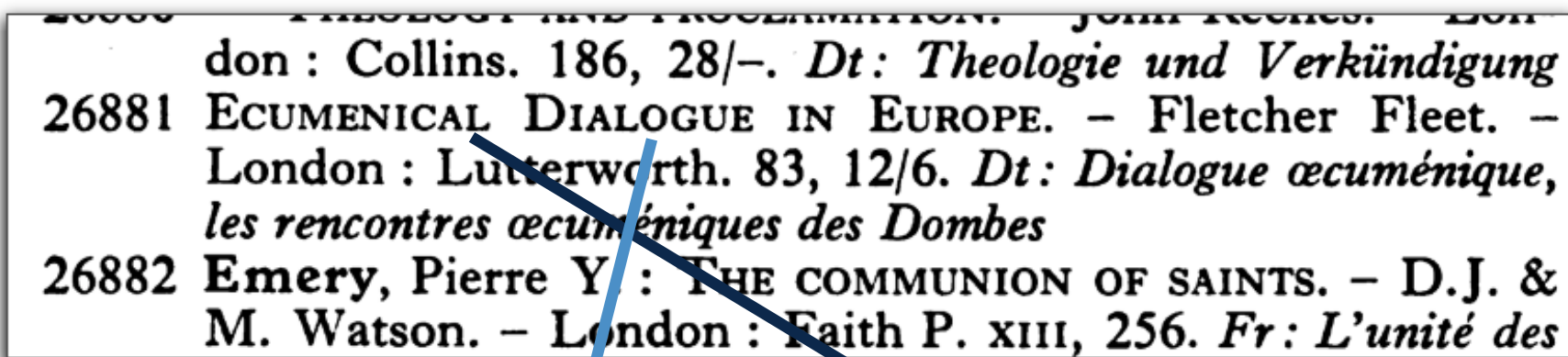
Unicode (1988) dont UTF-8
plus de 100 000 caractères
(dont > 3 000 émojis)
de 1 jusqu'à 6 octets

Des documents « images »

26866-26914

Royaume-Uni-United Kingdom

- 26866 Boss, Medard : A PSYCHIATRIST DISCOVERS INDIA. - Henry A. Frey. - London : Wolff, 1965. 192, 31/6. Dt: *Indienfahrt eines Psychiaters*
- 26867 Boutrais, Cyprien Burns & Oates. VII
- 26868 Buys, J. : CHRIST London : Chapmar du monde
- 26869 Carrier, Hervé : T - Arthur J. Arriert. 1965. 335, 30/-. *Fr: religieuse*
- 26870 Congar, Yves : I Loretz. - London *Chrétiens en dialogue*
- 26871 JESUS CHRIST. - 223, 25/-. *Fr: Jésus*
- 26872 TRADITION ANI Thomas Rainborou 90/-. *Fr: La traditi*
- 26873 Cristiani, Léon : London : Burns & Oates
- 26874 Daniel-Rops, Hen 1870-1939. - John 48/-. *Fr: L'Eglise a*
- 26875 Desbuquois, Gust bury Wells : Fowle
- 26876 Dournes, Jacques : - London : G. Chapman. 205, 50/-. *Fr: Dieu aime les païens*
- 26877 Drze, A. : LIVING IN CHRIST, liturgy and sacraments. - F.M. Gale & Jennifer Nicholas. - London : G. Chapman. 209, 8/6. *Fr: Jésus Christ*
- 26878 Ebeling, Gerhard : THE L - James W. Leitch. - Lo Dt: *Vom Gebet : Predigten*
- 26879 THE NATURE OF FAITH. - Collins. 191, 7/6. Dt: *Glaubers*
- 26880 THEOLOGY AND PROCLAI don : Collins. 186, 28/-. I
- 26881 ECUMENICAL DIALOGUE IN London : Lutterworth. 83, *les rencontres acuméniques d*
- 26882 Emery, Pierre Y. : THE c M. Watson. - London : F: *croissants au ciel et sur la t*
- 26883 Floristan, Casiano : THE P John F. Byrne. - London 11/6. *Esp: La parroquia, e*
- 26884 Geisslmann, Josef R. : W.J. O'Hara. - London : *Die Heilige Schrift und die*
- 26885 Häring, Bernhard : THE N London : Burns & Oates, I *Auftrag der Sakramente*
- 26886 Hérís, Charles V. : SPIRIT - London : Herder, 1965. *l'amour*
- 26887 Jaeger, Lorenz *Erzbischof* ECUMENISM : THE COUNCI London : G. Chapman, I *Konzildekret über den Ökum*
- 26888 Jeanne d'Arc, *Sœur* : V Martin Murphy. - London *Les religieuses dans l'Eglise*
- 26889 Jungmann, Josef A. : THE L. Batley. - London : Bu... *Das Eucharistische Hochgebet*
- 26890 LITURGICAL RENEWAL IN RETROSPECT AND PROSPECT. - Clifford Howell. - London : Burns & Oates, 1965. 45, 4/6. Dt: *Liturgische Erneuerung*
- 26891 THE LITURGY OF THE WORD. - H.E. Winstone. - London : Burns & Oates. 82, 8/6. Dt: *Wortgottesdienst im Lichte von*
- 26901 LITURGY IN DEVELOPMENT. - H.J.J. Vaughan. - London : Sheed & Ward, 1965. IX, 187, 12/6. *Ned: Medienst in*



```

2300 don</span> <span class='ocrx_word' title='bbox 301 279 307 285'>:</span> <span class='ocrx_word' title='bbox 335 226 341 232'>Collins.</span> <span
class='ocrx_word' title='bbox 514 226 520 232'>186,</span> <span class='ocrx_word' title='bbox 616 226 622 232'>28/</span>.</span> <span class='ocrx_word' title='bbox
736 226 742 232'>Dt:</span> <span class='ocrx_word' title='bbox 835 226 841 232'>Theologie</span> <span class='ocrx_word' title='bbox 935 226 941 232'>und</span>
<span class='ocrx_word' title='bbox 1035 226 1041 232'>Verkündigung</span>
26881 ECUMENICAL DIALOGUE IN EUROPE. - Fletcher Fleet. -
London : Lutterworth. 83, 12/6. Dt: Dialogue œcuménique,
les rencontres œcuméniques des Dombes
26882 Emery, Pierre Y. : THE COMMUNION OF SAINTS. - D.J. &
M. Watson. - London : Faith P. XIII, 256. Fr: L'unité des
<p class='ocr_par' title='bbox 68 2315 163 2406' style='font-size:8pt;font-family:"Times New Roman";font-style:normal'><span class='ocr_line' title='bbox 68 2315
1613 2350'><span class='ocrx_word' title='bbox 68 2315 179 2349'>26881</span> <span class='ocrx_word' title='bbox 211 2315 248 2348'>Ecumenical</span> <span
class='ocrx_word' title='bbox 491 2315 528 2349'>Dialogue</span> <span class='ocrx_word' title='bbox 721 2315 758 2348'>in</span> <span class='ocrx_word' title='bbox
793 2315 830 2349'>Europe.</span> <span class='ocrx_word' title='bbox 984 2315 1021 2339'>.</span> <span class='ocrx_word' title='bbox 1038 2315 1075 2317'>83,</span>
<span class='ocrx_word' title='bbox 1138 2315 1175 2317'>12/6.</span> <span class='ocrx_word' title='bbox 1238 2315 1275 2317'>Dt:</span> <span class='ocrx_word' title='bbox 1338 2315 1375 2317'>Dialogue
œcuménique,</span> <span class='ocrx_word' title='bbox 1438 2315 1475 2317'>les
rencontres œcuméniques des Dombes</span>
</p>
<p class='ocr_par' title='bbox 210 2410 247 2452' style='font-size:8pt;font-family:"Times New Roman";font-style:italic'><span class='ocr_line' title='bbox 210 2410
247 2452'><span class='ocrx_word' title='bbox 210 2410 256 2443'>les</span> <span class='ocrx_word' title='bbox 271 2410 308 2444'>rencontres</span> <span
class='ocrx_word' title='bbox 376 2410 413 2452'>œcuméniques</span> <span class='ocrx_word' title='bbox 473 2410 510 2445'>des</span> <span class='ocrx_word'
title='bbox 528 2410 565 2445'>Dombes</span>
</p>

```

Des balises orientées affichage

BACHELARD, Gaston

1975 *La formation de l'esprit scientifique* (Paris, Librairie philosophique J. Vrin)
[1^{re} éd. 1938].

BARTHES, Roland

1957 *Mythologies* (Paris, Le Seuil).

BENVENISTE, Émile

1966 De la subjectivité dans le langage, in É. Benveniste (dir.), *Problèmes de linguistique générale*, I (Paris, Gallimard) : 258-266.

1974 La forme et le sens dans le langage, in É. Benveniste (dir.), *Problèmes de linguistique générale*, II (Paris, Gallimard) : 215-240.

```
<div type="bibliography">
  <listBibl>
    <bibl><hi rend="bold">R</hi><hi rend="small-caps bold">mossy</hi><hi rend="bold">, Ruth</hi></hi>1997 <hi rend="
    <bibl><hi rend="bold">R</hi><hi rend="small-caps bold">ron</hi><hi rend="bold">, Raymond</hi></hi>1967 <hi rend="
    <bibl><hi rend="bold">R</hi><hi rend="small-caps bold">sséo</hi><hi rend="bold">, Henriette</hi></hi>1974 Le tr
    <bibl><hi rend="bold">B</hi><hi rend="small-caps bold">achelard</hi><hi rend="bold">, Gaston</hi></hi>1975 <hi
    <bibl><hi rend="bold">B</hi><hi rend="small-caps bold">arthes</hi><hi rend="bold">, Roland</hi></hi>1957 <hi re
    <bibl><hi rend="bold">B</hi><hi rend="small-caps bold">enveniste</hi><hi rend="bold">, Émile</hi></hi>1966 De l
    <bibl><hi rend="bold">C</hi><hi rend="small-caps bold">oupry</hi><hi rend="bold">, François</hi></hi>1999 <hi r
    <bibl><hi rend="bold">C</hi><hi rend="small-caps bold">ruz</hi><hi rend="bold"> S</hi><hi rend="small-caps bold
    <bibl><hi rend="bold">D</hi><hi rend="small-caps bold">escola</hi><hi rend="bold">, Philippe</hi></hi>1994 Rétr
    <bibl><hi rend="bold">D</hi><hi rend="small-caps bold">ubois</hi><hi rend="bold">, Jacques</hi></hi>1992 <hi re
    <bibl>
      <hi xml:lang="en" rend="italic bold">Etnoloska sticisca</hi>
      <hi xml:lang="en" rend="italic bold">
        <lb/>
      </hi>
      <hi xml:lang="en">1997 </hi>
      <hi xml:lang="en" rend="italic">Prejudices and stereotypes in the social sciences and humanities</hi>
      <hi xml:lang="en">, 5 et 7 (Ljubljana, Jezernik, Bozidrz Ed.).</hi>
    </bibl>
    <bibl><hi rend="bold">F</hi><hi rend="small-caps bold">erney</hi><hi rend="bold">, Alice</hi></hi>1997 <hi rend="
    <bibl><hi rend="bold">G</hi><hi rend="small-caps bold">eertz</hi><hi rend="bold">, Clifford</hi></hi>1986 Comme
    <bibl>
      <hi xml:lang="en" rend="bold">G</hi>
      <hi xml:lang="en" rend="small-caps bold">ilman</hi>
      <hi xml:lang="en" rend="bold">, Sander L.</hi></hi>
      <hi xml:lang="en">1985 </hi>
      <hi xml:lang="en" rend="italic">Difference and pathology, stereotypes of sexuality, race and madness</hi>
      <hi xml:lang="en"> (Ithaca, Cornell University Press).</hi>
    </bibl>
  </listBibl>
</div>
```


et des annotations sémantiques

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:ns1="http://standoff.proposal"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="https://xml-schema.delivery.istex.fr/formats/tei-istex.xsd">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title level="a" type="main" xml:lang="en">The design and realization of CoViD: a system for
          design</title>
      </titleStmt>
      <publicationStmt>
        <authority>ISTEX</authority>
        <publisher ref="https://scientific-publisher.data.istex.fr/ark:/67375/H02-SWLMH5L1-1">Springer</publisher>
        <pubPlace>London</pubPlace>
        <availability>
          <licence>Springer-Verlag London Limited</licence>
          <p scheme="https://loaded-corporus.data.istex.fr/ark:/67375/XBH-3XSW68JL-F">springer</p>
        </availability>
        <date type="published" when="2006">2006</date>
      </publicationStmt>
      <notesStmt>
        <note type="content-type"
          subtype="research-article"
          source="OriginalPaper"
          scheme="https://content-type.data.istex.fr/ark:/67375/XTP-1JC4F85T-7">research-article</note>
        <note type="publication-type"
          subtype="journal"
          scheme="https://publication-type.data.istex.fr/ark:/67375/JMC-5WTPMB5N-F">journal</note>
      </notesStmt>
      <sourceDesc>
        <biblStruct>
          <analytic>
            <title level="a" type="main" xml:lang="en">The design and realization of CoViD: a syst
              design</title>
            <author role="corresp">
              <persName>
                <forename type="first">Wolfgang</forename>
                <surname>Stuerzlinger</surname>
              </persName>
              <affiliation>
                <orgName type="institution">York University</orgName>
                <address>
                  <settlement>Toronto</settlement>
                  <country key="CA" xml:lang="en">CANADA</country>
                </address>
              </affiliation>
            </author>
          </analytic>
        </biblStruct>
      </sourceDesc>
    </fileDesc>
  </teiHeader>

```



VS

Abstract Many important decisions in the design process are made during fairly early on, after designers have presented initial concepts. In many domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today's systems and tools do not support this, as computer systems typically support only a single user and computer-aided design tools require significant training. This paper presents the design of a new system to facilitate a collaborative 3D design process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D design systems. Then, we introduce the new system, which consists of two main parts. The first part is an easy-to-use conceptual 3D design tool that can be used productively even by naive users. The tool provides novel interaction techniques that support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rapid creation and modification of 3D models. The second part is a novel infrastructure for collaborative work, which of a semi-immersive

W. Stuerzlinger (& amp;) L. Zaman A. Pavlovych
York University, Toronto, Canada
URL: <http://www.cs.yorku.ca/~wolfgang>
URL: <http://www.cs.yorku.ca/~zaman>
URL: <http://www.cs.yorku.ca/~andriyp>
J.-Y. Oh
University of Arizona, Tucson, AZ, USA
e-mail: jyoh@optics.arizona.edu

setup. It is designed to support multiple users working together. This infrastructure also includes novel pointing devices that work both as a stylus and a remote pointing device. collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present against the guidelines for collaborative 3D design. Finally, we present res which asked naive users to collaborate in a 3D design task on the new system.
Keywords Collaborative design 3D design
Collaborative virtual reality

<https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>

De multiples formats autres que XML

```
"jour";"nomReg";"numReg";"incid_rea"  
2020-03-19;"Auvergne-Rhône-Alpes";84;44  
2020-03-19;"Bourgogne-Franche-Comté";27;33  
2020-03-19;"Bretagne";53;8  
2020-03-19;"Centre-Val de Loire";24;6  
2020-03-19;"Corse";94;11  
2020-03-19;"Grand-Est";44;69  
2020-03-19;"Guadeloupe";1;0  
2020-03-19;"Guyane";3;0  
2020-03-19;"Hauts-de-France";32;37  
2020-03-19;"Ile-de-France";11;151  
2020-03-19;"La Réunion";4;0  
2020-03-19;"Martinique";2;0  
2020-03-19;"Mayotte";6;0  
2020-03-19;"Normandie";28;7  
2020-03-19;"Nouvelle-Aquitaine";75;7  
2020-03-19;"Occitanie";76;29  
2020-03-19;"Pays de la Loire";52;11  
2020-03-19;"Provence-Alpes-Côte d'Azur";93;25  
2020-03-20;"Auvergne-Rhône-Alpes";84;16  
2020-03-20;"Bourgogne-Franche-Comté";27;9  
2020-03-20;"Bretagne";53;2  
2020-03-20;"Centre-Val de Loire";24;4  
2020-03-20;"Corse";94;0  
2020-03-20;"Grand-Est";44;45
```

CSV

https://en.wikipedia.org/wiki/Comma-separated_values

```
"header": {  
  "title": "The JSON example",  
  "descriptionText": "This is some title text."  
},  
"content": {  
  "title": "The content example text",  
  "elements": [  
    {  
      "title": "The first element",  
      "mainText": "First element main text",  
      "additionalText": "First element additional te  
    },  
    {  
      "title": "The second element",  
      "mainText": "Second element main text",  
      "additionalText": "Second element additional
```

JSON et ses variantes

<https://en.wikipedia.org/wiki/JSON>

```
---  
receipt:    Oz-Ware Purchase Invoice  
date:      2012-08-06  
customer:  
  first_name: Dorothy  
  family_name: Gale  
  
items:  
- part_no:  A4786  
  descrip:  Water Bucket (Filled)  
  price:    1.47  
  quantity: 4  
  
- part_no:  E1628  
  descrip:  High Heeled "Ruby" Slippers  
  size:     8  
  price:    133.7  
  quantity: 1  
  
bill-to:   &id001  
  street:  |  
           123 Tornado Alley  
           Suite 16  
  city:    East Centerville  
  state:   KS  
  
ship-to:   *id001  
  
specialDelivery: >  
  Follow the Yellow Brick
```

YAML

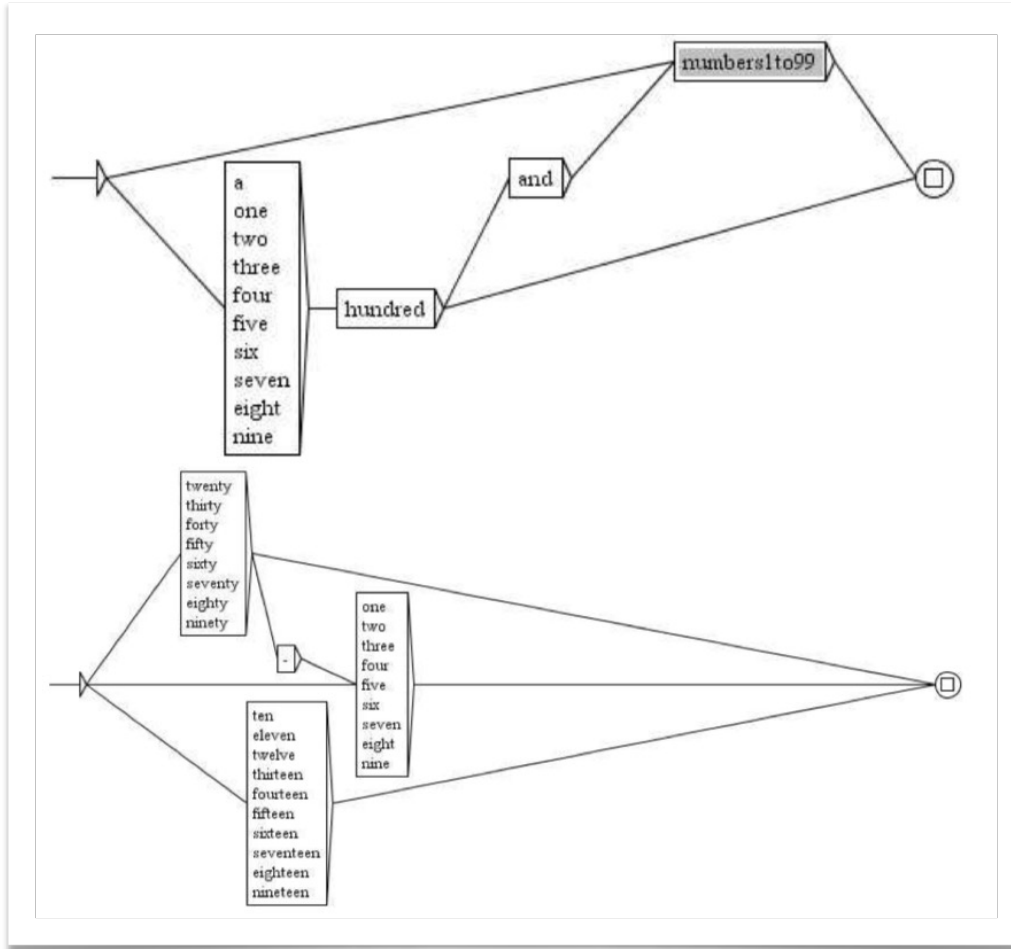
Yet Another Markup Language / YAML Ain't Markup Language

<https://en.wikipedia.org/wiki/YAML>

Des méthodes informatiques et des ressources linguistiques

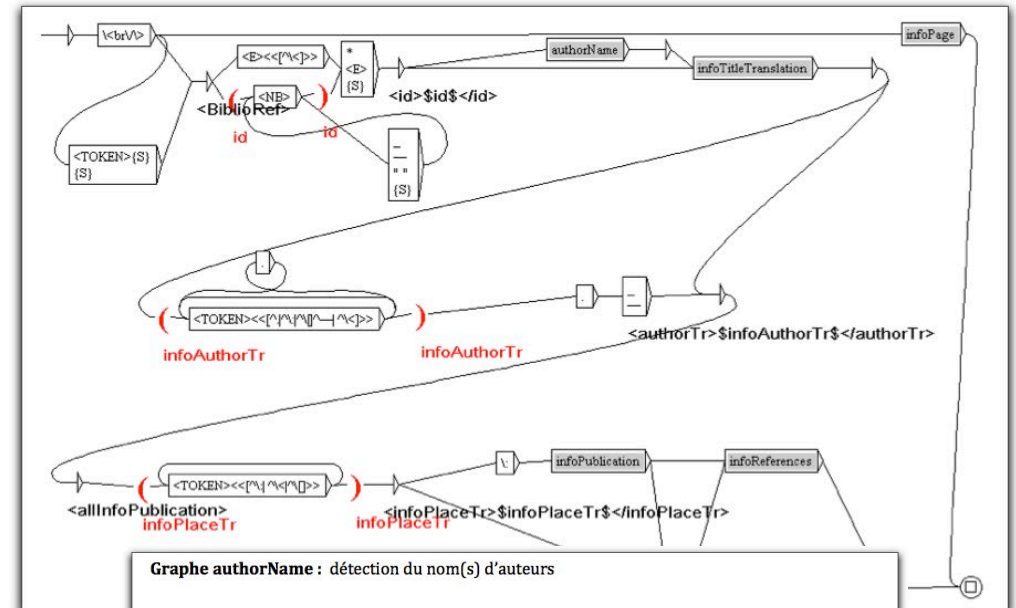
04.06.19

Des symboles et des automates

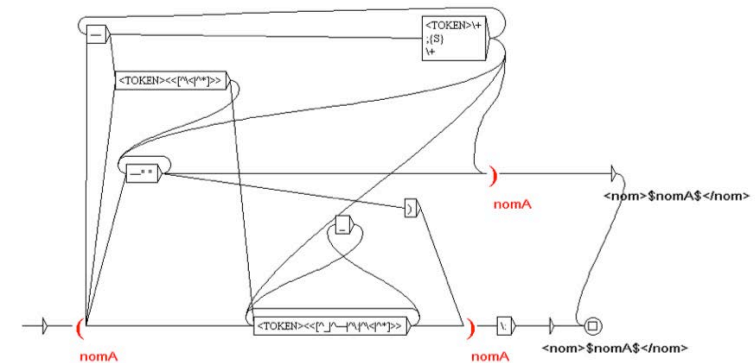


Eric Laporte. Symbolic Natural Language Processing. Lothaire. Applied Combinatorics on Words, Cambridge University Press, pp.164-209, 2005. hal-00145253

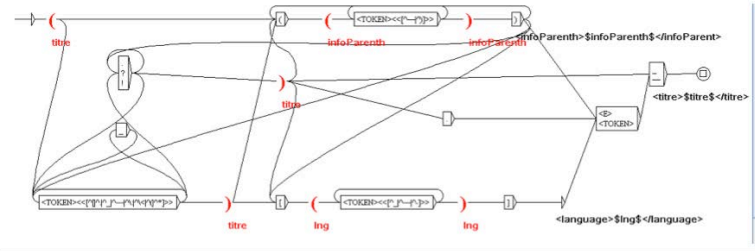
<https://hal.archives-ouvertes.fr/hal-00145253/document>



Grappe authorName : détection du nom(s) d'auteurs

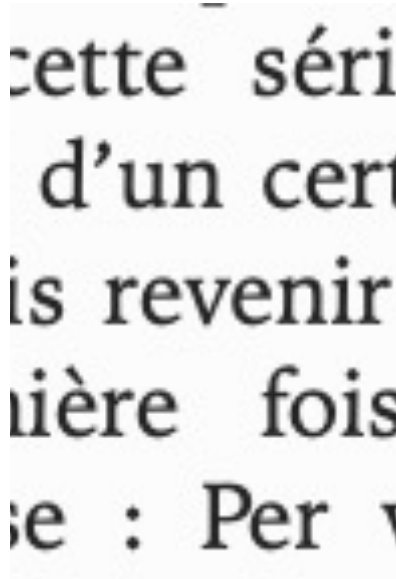


Grappe infoTitleTranslation : détection du titre de la traduction



Etude pour l'UNESCO (Univ. Avignon, Open Edition 2011)

Symboles, vecteurs et modèles de langue

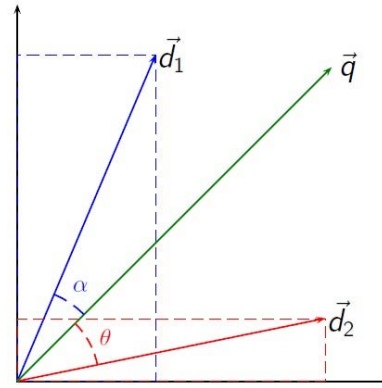


Un mot = une forme

$$Z_{\text{score}}(t_{ij}) = \frac{\text{tfr}_{ij} - \text{mean}_i}{\text{sdi}}$$

positive	Z_score	negative	Z_score	Neutral	Z_score
Love	14.31	Not	13.99	Httpbit	6.44
Good	14.01	Fuck	12.97	Httpfb	4.56
Happy	12.30	Don't	10.97	Httpbnd	3.78
Great	11.10	Shit	8.99	Intern	3.58
Excite	10.35	Bad	8.40	Nov	3.45
Best	9.24	Hate	8.29	Httpdlvr	3.40
Thank	9.21	Sad	8.28	Open	3.30
Hope	8.24	Sorry	8.11	Live	3.28
Cant	8.10	Cancel	7.53	Cloud	3.28
Wait	8.05	stupid	6.83	begin	3.17

Table1. The first ten terms having the highest Z_score in each class



Des vecteurs de mots

$$\text{Similarité}(d_1, d_2) \approx \vec{d}_1 \cdot \vec{d}_2$$

$$\text{Similarité}(d_1, d_2) \approx \cos(\vec{d}_1, \vec{d}_2)$$

Mot	Probabilité

Des modèles de langues

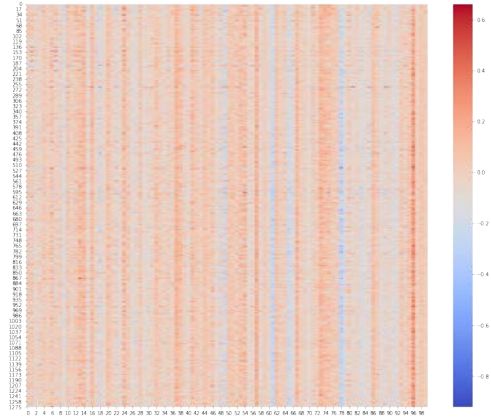
the cat sat on the mat $P(w_1)$
 the **cat** sat on the mat $P(w_2|w_1)$
 the cat **sat** on the mat $P(w_3|w_2, w_1)$
 the cat sat **on** the mat $P(w_4|w_3, w_2, w_1)$
 the cat sat on **the** mat $P(w_5|w_4, w_3, w_2, w_1)$
 the cat sat on the **mat** $P(w_6|w_5, w_4, w_3, w_2, w_1)$

Slide Credit: Piotr Mirowski

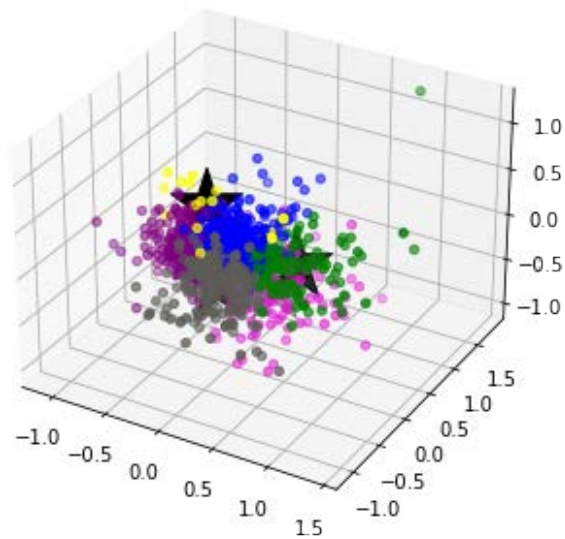
$$P(\text{Classe} | (w_1, w_2, \dots, w_{T-1}, w_T))$$

et règle de Bayes

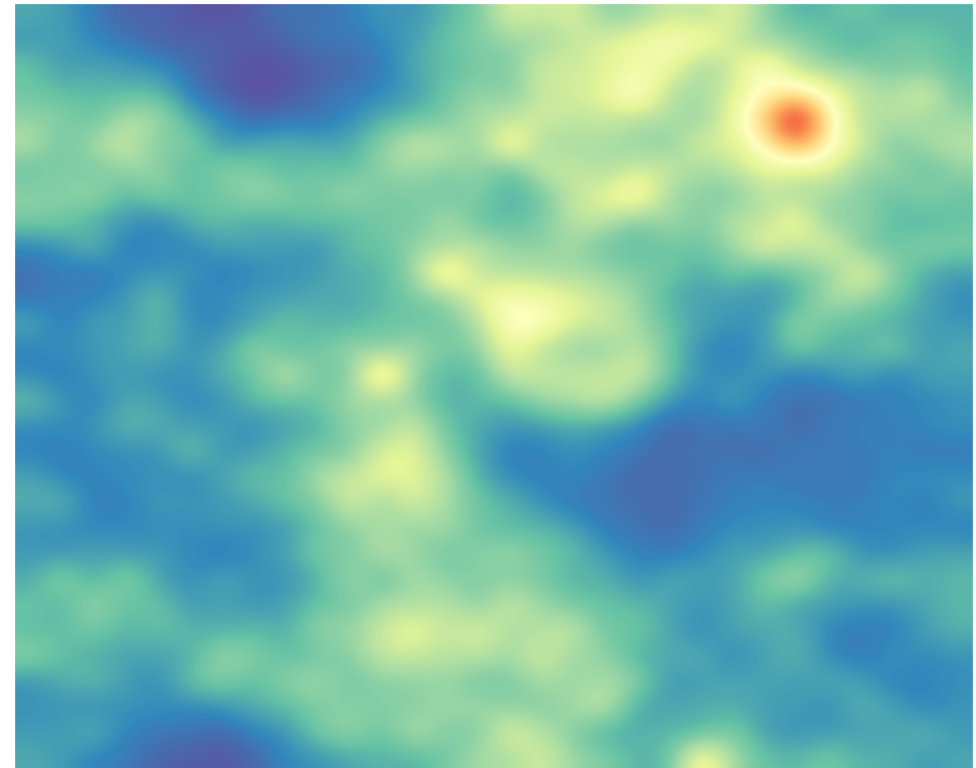
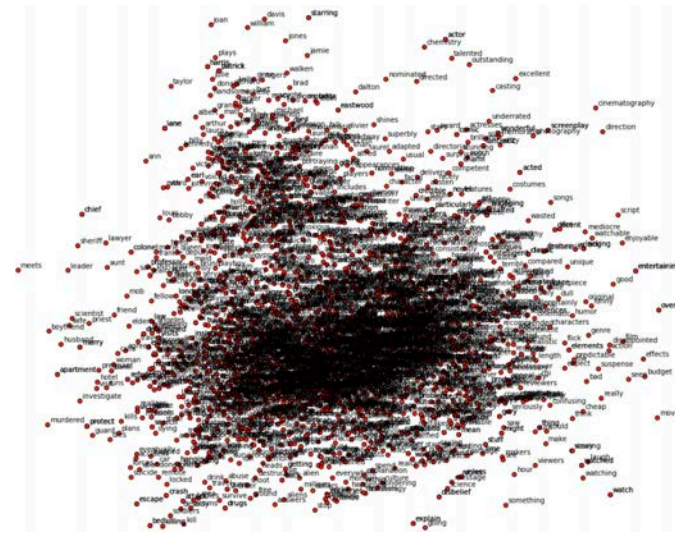
et des représentations



Vectorisation dans des espaces continus (Doc2Vec) par plongements lexicaux



Analyse en Composantes Principales




Cartes auto-organisées (SOM)

Un exemple de modèle de langue

Language	#Volumes	#Tokens
English	4,541,627	468,491,999,592
Spanish	854,649	83,967,471,303
French	792,118	102,174,681,393
German	657,991	64,784,628,286
Russian	591,310	67,137,666,353
Italian	305,763	40,288,810,817
Chinese	302,652	26,859,461,025
Hebrew	70,636	8,172,543,728

Table 1: Number of volumes and tokens for each language in our corpus. The total collection contains more than 6% of all books ever published.



The Google Books Ngram Viewer is optimized for quick inquiries into the usage of small sets of phrases. If you're interested prefer to download a portion of the corpora yourself. Or all of it, if you have the bandwidth and space. We're happy to oblige.

These datasets were generated in February 2020 (version 3), July 2012 (Version 2) and July 2009 (Version 1); we will update versions will have distinct and persistent version identifiers (20200217, 20120701 and 20090715 for the current sets).

Each of the numbered links below will directly download a fragment of the corpus. In Version 2 the ngrams are grouped alpha Version 1 the ngrams are partitioned into files of equal size. In addition, for each corpus we provide a file named `total_count` that make up the corpus. This file is useful for computing the relative frequencies of ngrams.

A summary of how the corpora were constructed can be found [here](#). We explain it in greater depth [here](#) (Version 2) and [here](#) appear over 40 times across the corpus. That's why the sum of the 1-gram occurrences in any given corpus is smaller than th

File format: Each of the files below is compressed *tab*-separated data. In Version 2 each line has the following format:

```
ngram TAB year TAB match_count TAB volume_count NEWLINE
```

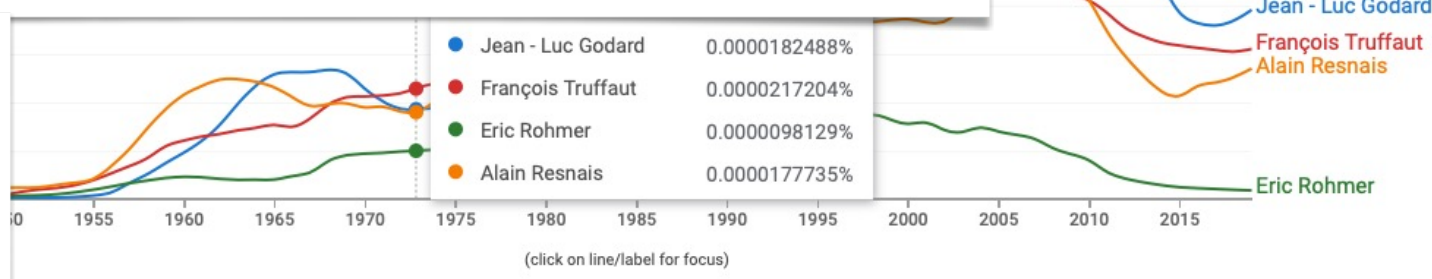
French

Version 20200217

- [1-grams](#)
- [2-grams](#)
- [3-grams](#)
- [4-grams](#)
- [5-grams](#)
- [Dependencies](#)

Version 20120701

[total_counts](#)



<https://books.google.com/ngrams>

Des modèles de langue « neuronaux »

ALBERT	DistilBERT	GPT-J
Auto Classes	DPR	GPT Neo
BART	ELECTRA	Hubert
BARThez	Encoder Decoder Models	Pegasus
BARTpho	FlauBERT	PhoBERT
BEiT	FNet	ProphetNet
BERT	FSMT	RAG
BERTweet	Funnel Transformer	Reformer
BertGeneration	HerBERT	RemBERT
BertJapanese	I-BERT	RetriBERT
BigBird	LayoutLM	RoBERTa
BigBirdPegasus	LayoutLMV2	RoFormer
Blenderbot	LayoutXLM	SegFormer
Blenderbot Small	LED	SEW
BORT	Longformer	SEW-D
ByT5	LUKE	Speech Encoder Decoder
CamemBERT	LXMERT	Speech2Text
CANINE	MarianMT	Speech2Text2
CLIP	M2M100	Splinter
ConvBERT	MBart and MBart-50	SqueezeBERT
CPM	MegatronBERT	T5
CTRL	MegatronGPT2	T5v1.1
DeBERTa	MobileBERT	TAPAS
DeBERTa-v2	MPNet	Transformer XL
DeiT	mT5	TrOCR
DETR	OpenAI GPT	UniSpeech
DialoGPT	OpenAI GPT2	UniSpeech-SAT



CamemBERT

A Tasty French Language Model

Facebook AI Research

Inria

ALMAnaCH

<https://camembert-model.fr>

FlauBERT

<https://github.com/getalp/Flaubert>



projector.tensorflow.org

word embedding explorer - Recherc... Embedding projector - visualization... wevi Vector explorers Word embedding demo

Embedding Projector

DATA

5 tensors found
Word2Vec 10K

Label by
word

Color by
No color map

Sphereize data

Load data Publish

Checkpoint: Demo datasets
Metadata: oss_data/word2vec_10000_200d_labels.tsv

T-SNE **PCA** CUSTOM

X Component #1 Y Component #2
Z Component #3

PCA is approximate.
Total variance described: 8.5%

Points: 10000 | Dimension: 200

Show All Data Isolate 5948 points Clear selection

Search by word

<https://projector.tensorflow.org>

BOOKMARKS (0)

projector.tensorflow.org

word embedding explorer - Recherc... Embedding projector - visualization... wevi Vector explorers Word embedding demo

Embedding Projector

Points: 10000 | Dimension: 200 | Selected 101 points

DATA

5 tensors found
Word2Vec 10K

Label by
word

Color by
No color map

Sphereize data

Load data Publish

Checkpoint: Demo datasets
Metadata: oss_data/word2vec_10000_200d_labels.tsv

Show All Data Isolate 101 points Clear selection

Search ee by word

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

open	0.578
freely	0.580
available	0.592
freedom	0.616
software	0.638
source	0.647
online	0.678
tools	0.686
independent	0.687
allow	0.689
content	0.690
complete	0.695
new	0.696
good	0.696
allowing	0.699
support	0.702
bound	0.706
access	0.709
create	0.710
licensing	0.711
thus	0.711
strong	0.718
download	0.719
developers	0.720
fair	0.722

BOOKMARKS (0)

T-SNE **PCA** CUSTOM

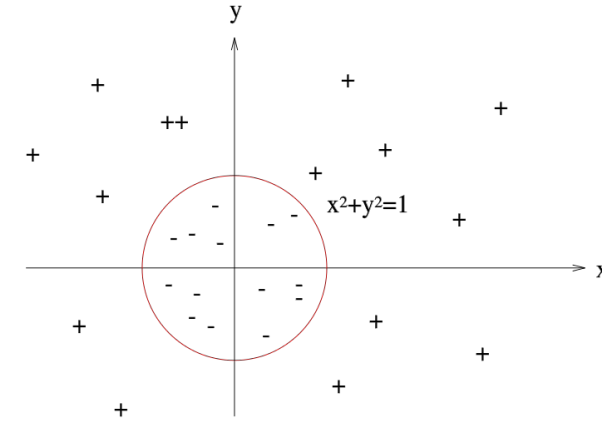
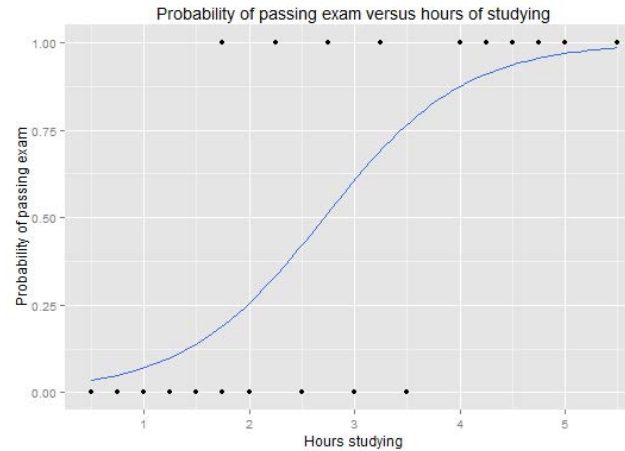
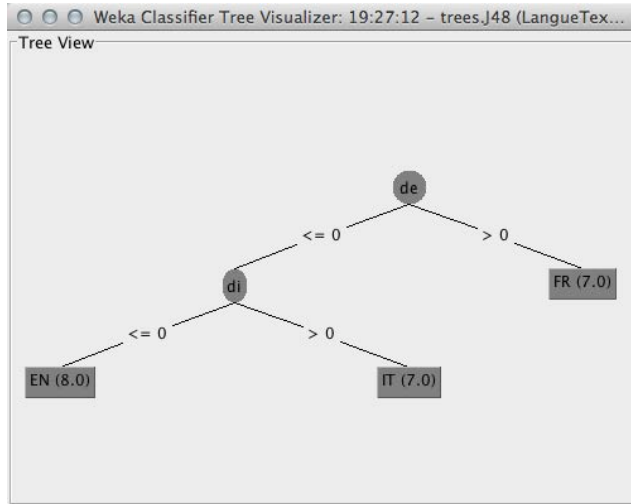
X Component #1 Y Component #2

Z Component #3

PCA is approximate.

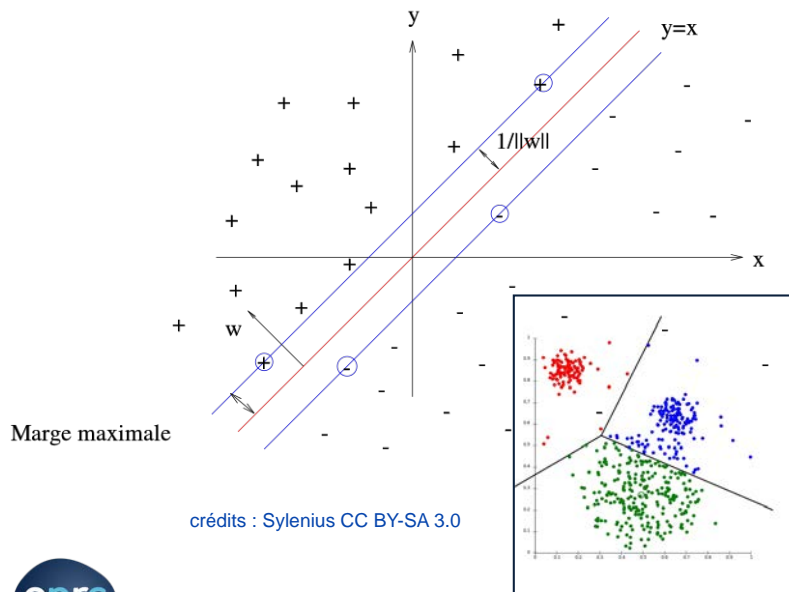
Total variance described: 8.5%.

Arbres et forêts, régressions, vastes marges, neurones...

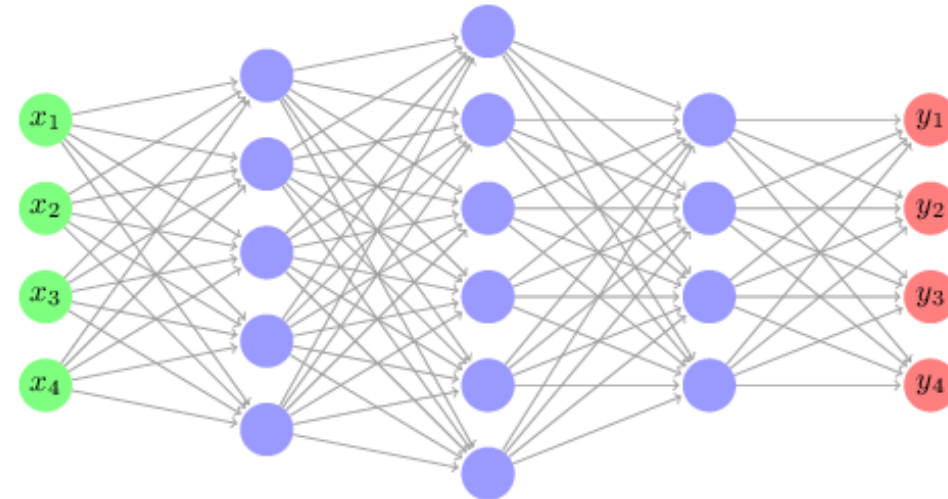


crédits : Michaelg2015 CC BY-SA 4.0

crédits : Sylenius CC BY-SA 3.0



crédits : Sylenius CC BY-SA 3.0



données observées

Sorties avec scores

Virtual Language Observatory (VLO)



The Virtual Language Observatory (VLO) provides a means of exploring language resources and tools. Its aim is to provide an easy to use interface, allowing for a uniform search and discovery process for a large number of resources from a wide variety of domains. Facets make it easy to explore and access available resources. A powerful query syntax makes it possible to carry out more targeted searches as well. It also makes it easy to review processing options for discovered resources via the Language Resource Switchboard, and to create virtual collections based on search results via the Virtual Collection Registry.

➔ [Go to the Virtual Language Observatory](https://www.clarin.eu/content/virtual-language-observatory-vlo)


The following list provides a few links for example selections and queries to start exploring:


- [Resources for spoken French](#)
- [Corpora with Polish content](#)
- [All records from the Language Bank of Finland](#)
- Searching for a general term: “slovenian news sentiment”
- Searching for a specific record or set of records: “Hamburg MapTask Corpus”


More information is available in the [VLO's integrated help page](#).



<https://www.clarin.eu/content/virtual-language-observatory-vlo>

 ORTOLANG


 Accueil

 Corpus


Outils et Ressources pour un Traitement Optimisé de la LANGUE


ORTOLANG est un équipement d'excellence validé dans le cadre des [investissements d'avenir](#). Son but est de proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés qui :

- permette, au travers d'une véritable mutualisation, à la recherche sur l'analyse, la modélisation et le traitement automatique de notre langue de se hisser au meilleur niveau international;
- facilite l'usage et le transfert des ressources et outils mis en place au sein des laboratoires publics vers les partenaires industriels, en particulier vers les PME qui souvent ne peuvent pas se permettre de développer de telles ressources et outils de traitement de la langue compte tenu de leurs coûts de réalisation;
- valorise le français et les langues de France à travers un partage des connaissances sur notre langue accumulées par les laboratoires publics.


 ORTOLANG est un service spécialisé pour la langue, complémentaire de l'offre générale proposée par [Huma-Num](#) (très grande infrastructure de recherche).


La charte d'ORTOLANG définit les modalités d'utilisation et de dépôt de ressources sur la plate-forme.


Vous pouvez [consulter la charte](#) ou la télécharger ( [fichier au format pdf](#))


 Lexiques


ORTOLANG bénéficie d'une aide de l'Etat au titre du programme « Investissements d'avenir » (ANR-11-EQPX-0032)
ORTOLANG ISSN 2417-7482


 Terminologies

 Outils

 Projets Intégrés

 Actualités

 Informations

 Producteurs

<https://www.ortolang.fr>



elra catalogue

1,387 language resources at your disposal

This is the new version of the ELRA Catalogue of Language Resources. If you would like to view the older version, [click here](#)



An increasing number of LRs in the various fields of Human Language Technology (see image on the left-hand side) are distributed on behalf of ELRA via its operational body ELDA, thanks to the con- the HLT community.
Our aim is to provide Language Resources, by means of this repository, so as to prevent researchers and developers from investing efforts to rebuild resources which already exist as well as help their resources.

Latest Resources

French-Vietnamese Dictionary

The French-Vietnamese Dictionary consists of 82,768 entries containing the following information: phonetics (using IPA), morphology, grammar, semantics, pragmatics and examples. All headwords are pronounced with true voice by native speakers. The dictionary is provided in XML format.

German-Vietnamese Dictionary

The German-Vietnamese Dictionary consists of 32,511 entries containing the following information: phonetics (using IPA), morphology, grammar, semantics, pragmatics and examples available only for the source language. Headword (in Vietnamese) has true voice by native speakers.

Vietnamese-French Dictionary

The Vietnamese-French Dictionary consists of 43,296 entries containing the following information: phonetics (using IPA), morphology, grammar, semantics, pragmatics and examples for source language only. The dictionary is provided in XML format.

Vietnamese-German Dictionary

The Vietnamese-German Dictionary consists of 42,793 entries containing the following information: phonetics (using IPA), morphology, grammar, semantics, pragmatics and examples available only for the source language.

Ema-Ion Manipuri Corpus (including word embedding and language model)

The Ema-Ion Manipuri Corpus consists of a set of resources for Manipuri language (locally known as Meiteilon) for the purpose of machine translation. The main source for these resources is the Sangai Express news website. The resources that constitute the present corpus are listed below: 1. EM Corpus, abbreviation of ...

NRC Emotion Lexicon - Revised version

The NRC Emotion Lexicon was originally built by Saif M. Mohammad and Peter D. Turney through crowdsourcing. The NRC was created in order to assist with emotion analysis as other emotion lexicons were smaller at the time. In order to be able to fix this problem, Saif crowdsourced a huge ...

<http://catalog.elra.info/en-us/>

voir aussi : <https://www ldc.upenn.edu>

Ressources « collaboratives »



Diko le dictionnaire d'associations lexicales
contributif et libre de JeuxDeMots

Chercher la forme >>>13:68887>29:80995

compter [sujet] élève Verbe infinitif, Chunk:

Informations diverses wiki polarité

Associations d'idées > compter > élève ■ moutons ■ mouton > doigts < compter >

Génériques H > compter >

Qui/quoi peut compter [sujet... ?] > élève

Qui/Que peut-on compter [sujet... ?] > doigts ■ jours ■ minutes ■ pièces ■ moutons ■ heures

Où peut-on compter [sujet... ?] > école > ■ collègue > ■ lycée ■ école primaire ■ salle de classe

Avec quoi peut-on compter [sujet... ?] > boulier >

compter [sujet... comme sujet] < élève

compter [sujet... comme prédicat] < compter >

Champs masqués vide

J'aime Soyez le premier de vos amis à indiquer que vous aimez ça.



Open Academic Graph

OpenAlex

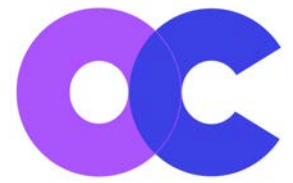
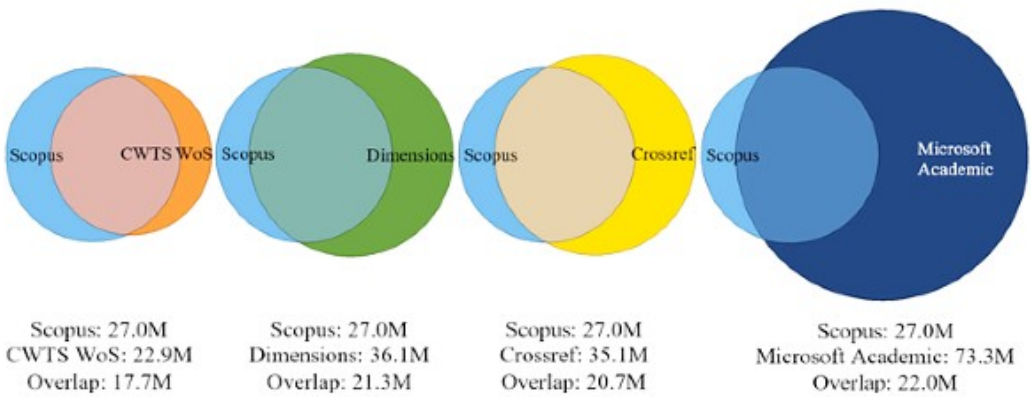
<https://openalex.org>

An open, comprehensive catalog of scholarly papers, authors, institutions, and more.

OpenAlex will launch in December 2021, as a drop-in replacement for [Microsoft Academic Graph](#). Learn more in our latest blog post, and join the mailing list to stay up-to-date.

Overview Publications

Open Academic Graph (OAG) is a large knowledge graph unifying two billion-scale academic graphs: [Microsoft Academic Graph](#) (MAG) and [AMiner](#). In mid 2017, we published OAG v1, which contains 166,192,182 papers from MAG and 154,771,162 papers from AMiner (see below) and generated 64,639,608 linking (matching) relations between the two graphs. This time, in OAG v2, author, venue and newer publication data and the corresponding matchings are available.



<https://opencitations.net>

Credit: Martijn Visser, Nees Jan van Eck and Ludo Waltman Quantitative Science Studies 2021; 2 (1): 20–41.

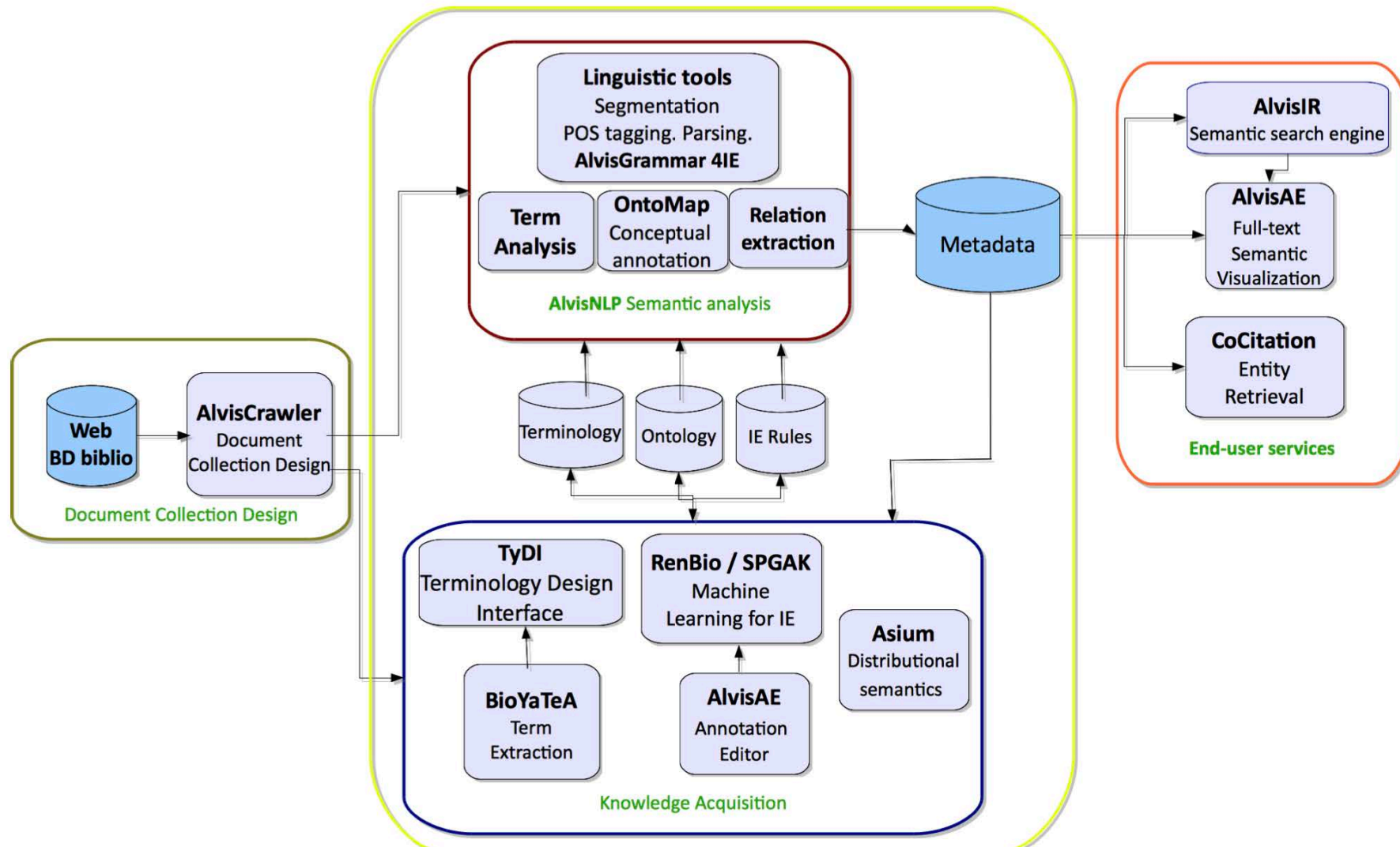
<https://www.natureindex.com/news-blog/microsoft-academic-graph-discontinued-whats-next>



Des services logiciels
orientés
« fouille de textes »



Plateforme Alvis





Info

Gargantext

A web platform to explore text-mining

Log in

Sign Up

Documentation

Some features may not work without a javascript optimized browser (Chromium for instance).



dépasser les frontières



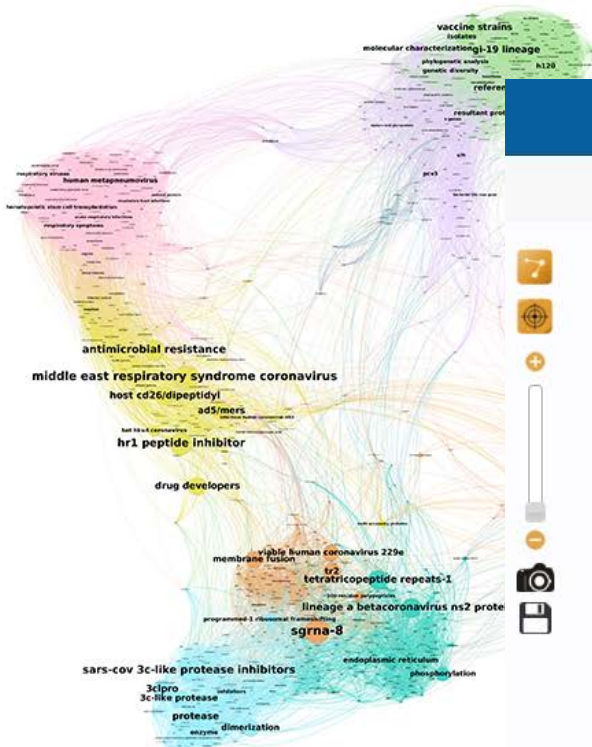
Add a Corpus ?

Type: Select a database below

Name: Select a database below

File: Europresse, Jstor [RIS], Pubmed [XML], Scopus [RIS], Web of Science [ISI], Zotero [RIS], CSV, IS-TeX, SCOAP [XML], REPEC [RIS]

Close Process this!



Documents Sources

Nodes: 18 Edges: 40 Label size: -1 Colors: Sizes: Selector size: 60

weight gain / significant decrease / doses / gastric intubation / high dose / mg/kg dose / developmental

Read More Delete

Neighbors

mice resorptions

bpa treatment

reproductive systems

Pubs (5/42)

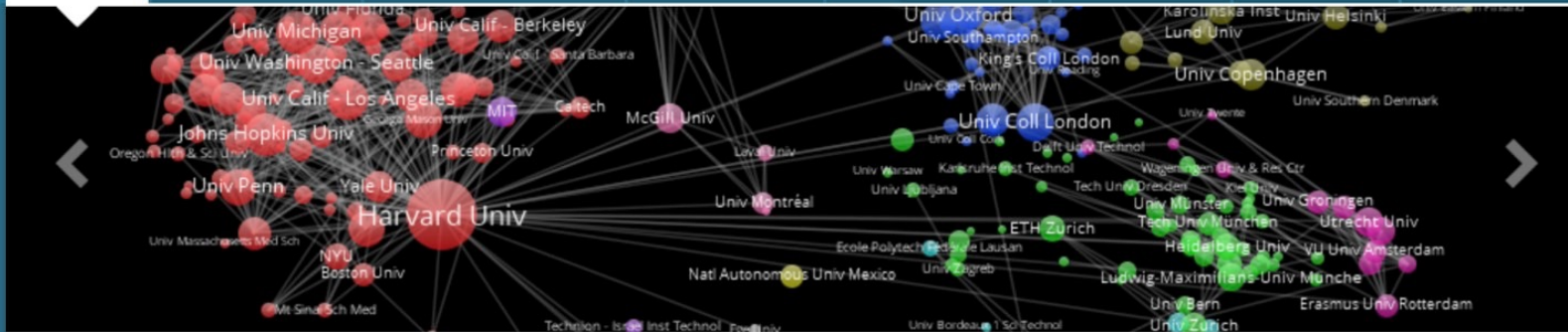
The Developmental Toxicity of Bisphenol A in Rats and Mice. . .
Published in Toxicological Sciences. . .
In 1987-01-01

The developmental toxicity of bisphenol A in rats and mice. . .
Published in Fundamental and Applied Toxicology. . .
In 1987-01-01

<https://gargantext.org>



<https://lejournal.cnrs.fr/articles/visualiser-la-recherche-sur-le-coronavirus-en-un-coup-doeil>



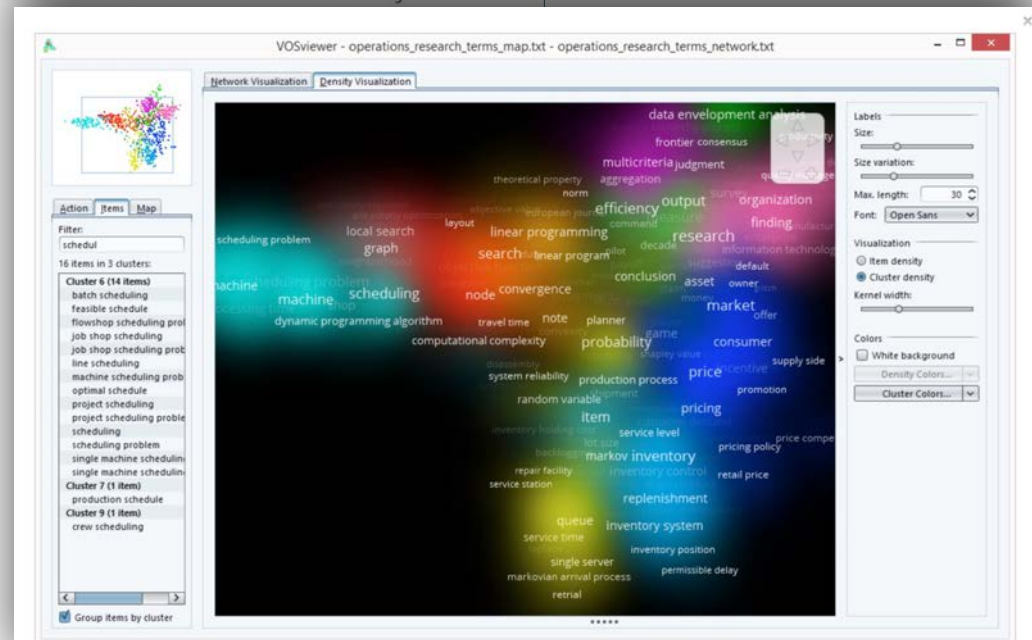
Welcome to VOSviewer

VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed based on co-citation, bibliographic coupling, or co-authorship relations. VOSviewer also offers text mining functionality that can be used to construct and visualize co-occurrence networks of important terms extracted from a body of scientific literature.

VOSviewer version 1.6.5

VOSviewer version 1.6.5 was released on September 28, 2016. Some of the improvements introduced in this version are listed below:

- **Overlay visualizations.** These popular visualizations have been made more prominently visible.
- **Maps based on bibliographic data.** Functionality for creating maps based on bibliographic data has been improved. Items can be filtered based on citation counts, and various types of overlay visualizations are supported.
- **Command line parameters.** Many



De nombreux acteurs industriels



Transformation digitale des entreprises - Xerox

À propos | Services | Produits | Fournitures | Support Client | Où acheter

Partager | Tweet | Lien | Courriel électronique | Imprimer | Plus

Analyse documentaire : l'arme secrète pour être en première ligne de la transformation numérique

Pour la plupart des entreprises, les ambitions numériques restent lettres mortes, comme le montre une enquête Xerox réalisée auprès de responsables IT. Le désir d'abandonner le papier pour



Elsevier R&D Solutions FOR PHARMA & LIFE SCIENCES

Pathway Studio® Fact Sheet

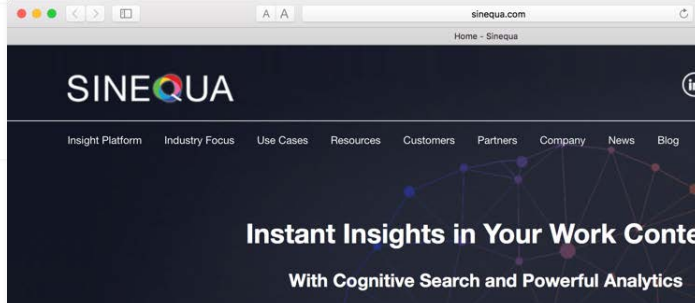


Linguamatics an IQVIA company

Technology Products Solu



AI Siblings: NLP and Machine Learning for Better Drug Discovery

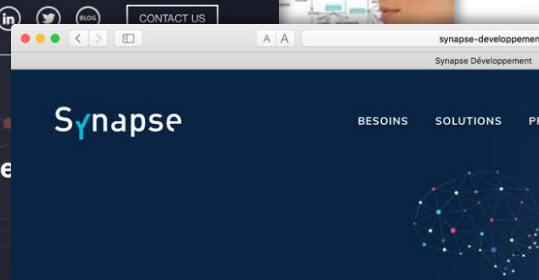


SINEQUA

Insight Platform | Industry Focus | Use Cases | Resources | Customers | Partners | Company | News | Blog

Instant Insights in Your Work Conte

With Cognitive Search and Powerful Analytics



Synapse

BESOINS SOLUTIONS PR

SYNAPSE DÉVELOPPEMENT



MONDECA

MAKING SENSE OF CONTENT

COMPRENDRE ET INTERPRETER

Comprendre le texte, l'image, la vidéo. Identifier ce qui est significatif. Catégoriser. Evaluer. Trouver des relations. Désambigüiser. Enrichir. Annoter. Gérer la connaissance pour nourrir les solutions d'intelligence artificielle: Classification automatique, bots, gestion d'alertes.



syllabs

Accueil | Immobilier | Nos offres d'emploi | Contact | EN | ES

Entrez dans le futur de la création de contenu

Syllabs propose des solutions automatisées de création de textes et d'optimisation de contenus. Notre approche unique qui conjugue expertise humaine et intelligence artificielle permet de répondre aux besoins d'information de tous vos publics, de développer votre trafic et d'optimiser votre stratégie SEO.



SYNAPSE DÉVELOPPEMENT

Les experts de l'Intelligence Artificielle appliquée au texte

- RÉINVENTER ma relation client
- VALORISER mes contenus
- OPTIMISER mes écrits

Nous utilisons les cookies pour vous offrir la meilleure expérience sur ce site. En continuant votre visite vous acceptez cette utilisation. En savoir plus

Devon Think

<https://www.devontechnologies.com/apps/devonthink>

The screenshot displays the DevonThink 3 application interface, which is used for organizing and searching through large volumes of text. The main window shows a list of documents with columns for Name, Created, Kind, and Size. A search bar at the top right allows for filtering documents based on various criteria.

On the right side, a sidebar displays a list of documents, including "Operative report 10/28/2007", "X-ray Chest 10/28/2007", "Critical Care Progress 10/28/2007", "Pathology from Surgery 10/30/2007 Report Date", "Labs 10/31/2007 Lipase High", "CT scan 10/31/2007", "Pathology report for 10/31/2007 surgery", "Operative report 11/1/2007", "Critical care progress note 11/1/2007", and "X-ray Chest 11/1/2007". A red arrow points to the search bar with the text: "Powerful Search: Limit to Current Selection, Expand to Encompass All, Open Databases".

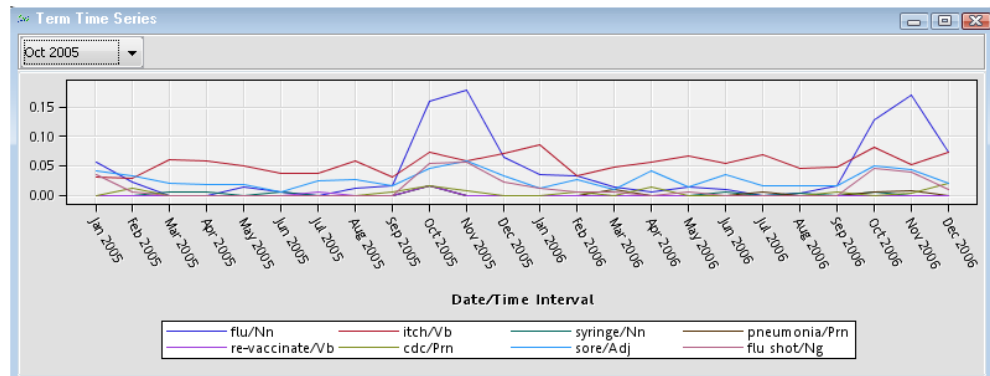
Below the main window, a smaller window displays a word index for the selected document. The word index shows the frequency of words and their context. A red arrow points to the word index with the text: "Word Index for Current Selection".

Another window on the right shows a list of documents related to the selected document, including "Pancreatic Necrosis and Pancr...", "Pathology report for 10/31/2007 surgery", "Pancreatic Necrosis and Pancr...: eMedicine Gastroenterology", "Operative report 10/28/2007", "Critical Care Progress 10/28/2007", "Operative Report 10/1/2007", "CT scan 10/31/2007", "Annotation: Def Supp Resp RFP.pdf", "f/u discuss ileostomy takedown 8/5/2008", "Pathology from Surgery 10/30/2007 Report Date", "CT scan abdomen 11/21/2007", and "2011 03 15 phonecall .rtf". A red arrow points to this list with the text: "Other Documents Closely Related According to Text Content".

The bottom window shows a document titled "CERAMICS" with the following text: "Ceramic materials commonly used in 3-D printing include those described in the following table." The table lists various ceramic materials and their properties, such as Alumina silica, Aluminum oxide, Zirconium oxide, and Tricalcium phosphate. A red arrow points to the table with the text: "Table of Imaging".

SAS® Text Miner

Mettez en évidence les informations dissimulées dans les données non structurées



Text Miner - Interactive

File Edit Tools View Window

SUMMARY

	NUMBER OF OCCURRENCES	VEHICLE/MAKE
REAR END COLLISION CAUSED SEAT BACKING TO COLLAPSE. *AK	0.0	TOYOTA TRUCK
WHILE WASHING TRUCK ACCIDENTLY RAISED HAND-OVER WINDSHIELD WIPER. REPLACEMENT (PYLON MOUNT) CUTTING HAND	1.0	TRUCK
VEHICLE SPRING BRAKES LOOKED UP AND VEHICLE SLID, RESULTING IN A COLLISION. TOOK VEHICLE TO THE DEALER. ALSO, EXPERIENCED SOME VERY ERRATIC VIBRATION WHILE IN TRAFFIC. PLEASE PROVIDE ANY FURTHER INFORMATION. *AK	0.0	TOYOTA TRUCK
WHILE TRAVELING 30-35 MPH CONSUMER WAS BLINDED BY ANOTHER VEHICLE'S FOG LIGHTS AND VEHICLE RAN INTO A TELEPHONE HEAD-ON, AIR BAGS DID NOT DEPLOY AT ANY TIME. MANUFACTURER HAS BEEN NOTIFIED. PLEASE PROVIDE FURTHER INFORMATION. *AK	0.0	TOYOTA TRUCK
WHILE PULLING VEHICLE IN PARK VEHICLE JUMPED OUT OF GEAR AND ACCELERATED FORWARD, RESULTING IN AN ACCIDENT. THIS OCCURRED SEVERAL TIMES. *AK	0.0	SUBARU TRUCK
FIRST REAR SEAL, REAR SEAL AND BEARING, BEARING AND AXLE, NEW AXLE PUT IN THEN FLEW OFF TRUCK AND CRASHED, 3 WEEKS LATER REAR SEALS AXLE AGAIN HELP	4.0	TOYOTA TRUCK
VEHICLE WAS INVOLVED IN TWO ACCIDENTS IN WHICH THERE WAS NO DEPLOYMENT OF DRIVERS OR PASSENGERS SIDE AIRBAGS. THE FINAL ACCIDENT WAS A DIRECT FRONTAL IMPACT AT 40 MPH IN WHICH DRIVER WAS HURT. CAUSE OF PROBLEM UNKNOWN. *AK	2.0	TOYOTA TRUCK
UPON ACCELERATION AND MAKING A LEFT HAND TURN, STEERING WHEEL STUCK, CAUSING POOR STEERING CONTROL. (CONSUMER HAS CONTACTED DEALER, DEALER HAS YET TO DETERMINE THE CAUSE. PLEASE PROVIDE ANY FURTHER DETAILS. *AK	0.0	TOYOTA TRUCK
WHILE TRAVELING 35 MPH TRUCK SIMPLY GOT CAUGHT BETWEEN TWO TREES ON SIDE OF ROAD. TRUCK THEN ROLLED TO RIGHT, AND OWNER TRIED TO MAINTAIN STABILITY. TRUCK WENT OFF THE ROAD, ROLLING OVER SEVERAL TIMES. SEAT BELT BECAME LOOSE, AND NO AIR BAG DEPLOYED. THIS IS A RENTAL VEHICLE. PLEASE DESCRIBE DETAILS. *AK	0.0	TOYOTA TRUCK
CONSUMER WAS INVOLVED IN A 35 MPH FRONTAL COLLISION IN WHICH THE DRIVER/PASSENGER'S AIR BAGS DID NOT DEPLOY. PLEASE GIVE ANY FURTHER DETAILS. *AK	0.0	PLYMOUTH TRUCK
WHILE STOPPING AT A TRAFFIC LIGHT VEHICLE WAS REAR ENDED AT 30 MPH, FORCED INTO ANOTHER VEHICLE HEAD-ON. UPON IMPACT, DRIVERS AND PASSENGER'S AIR BAGS DID NOT DEPLOY. *AK THE ACCIDENT CAUSED DAMAGE TO THE FRONT END OF THE VEHICLE AND INJURIES TO THE DRIVER. *YN	1.0	TOYOTA TRUCK

Clusters

#	Descriptive Terms	Freq	Percentage	PMS Stat
1	+ depth, + summary, + above, + below, + left, + deployment, + air	275	0.0303642467941	0.000466558
2	+ left, + number, + left number, + size, + fire, + rear, + firestone, + control	309	0.05466554043667	0.057529618
3	+ acceleration, + sudden, + accident, + sudden acceleration, + causing accident, + cause, + reverse, + attorney	111	0.0155995502746	0.04995817
4	+ tail, + left, + brake, + wheel, + side, + brake, + rear, + cone	463	0.06096464005789	0.00037296
5	+ fire, + fuel, + causing accident, + result, + airlock, + fire, + fire, + accident	400	0.06021136460384	0.001622413
6	+ out of, + down, + part, + into, + jump, + park, + gear, + rot	605	0.0801367337474	0.0571010590
7	+ air, + slip, + brake, + left, + fire, + left, + when, + drive	2430	0.34031016426032	0.10741591
8	+ break, + left, + passenger, + injury, + seat, + during, + impact, + rear	503	0.07060577952726	0.009990240
9	+ apply, + foot, + result, + brake, + stop, + when, + failure, + tail	716	0.10061032400763	0.034832675
10	+ rot, + brake, + air, + depth, + collision, + side, + left, + fire	1173	0.16483079763972	0.094722741

Concept Linking

Enhance your feedback insights with text analytics

Keatext is an AI-driven text analytics platform that instantly processes your unstructured feedback, enabling you to dive into the data and find key insights right away.

Take your feedback analysis to the next level with automatic comment grouping, opinion and sentiment analysis, advanced data visualization, and powerful trend detection.



Multichannel analysis

Upload and get a global view on your feedback data from reviews, emails, surveys, help desk tickets, and call centre logs.

Opinion and sentiment analysis

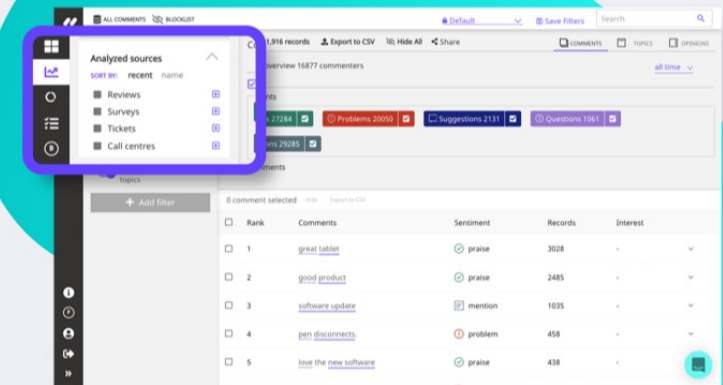
Instantly access the Praises, Problems, Suggestions, Questions, and Mentions in your feedback with accurate data categorization.

Multilingual processing

Analyze feedback with native proficiency in English and French, and top-of-the-line AI that translates 50+ other languages to English.

Advanced data filtering

Control and segment your data view with custom filters based on your metadata categories and relevant keywords.



Depth for Data Scientists, Simplified for Everyone Else

Depth

- 1,500+ native algorithms, data prep & data science functions
- Support for any 3rd party ML libraries
- Notebooks & integration with custom Python & R
- Advanced analytics & powerful platform services

Augmented data science - guided & automated:

- Data cleansing & transformation
- Algorithm selection & validation
- Visual, intuitive model operations

Simplified

- Solution accelerators (pre-canned use case templates)
- Comprehensive tutorials
- Abstract complexity
- Self-paced online certification by persona
- Full automation where desired

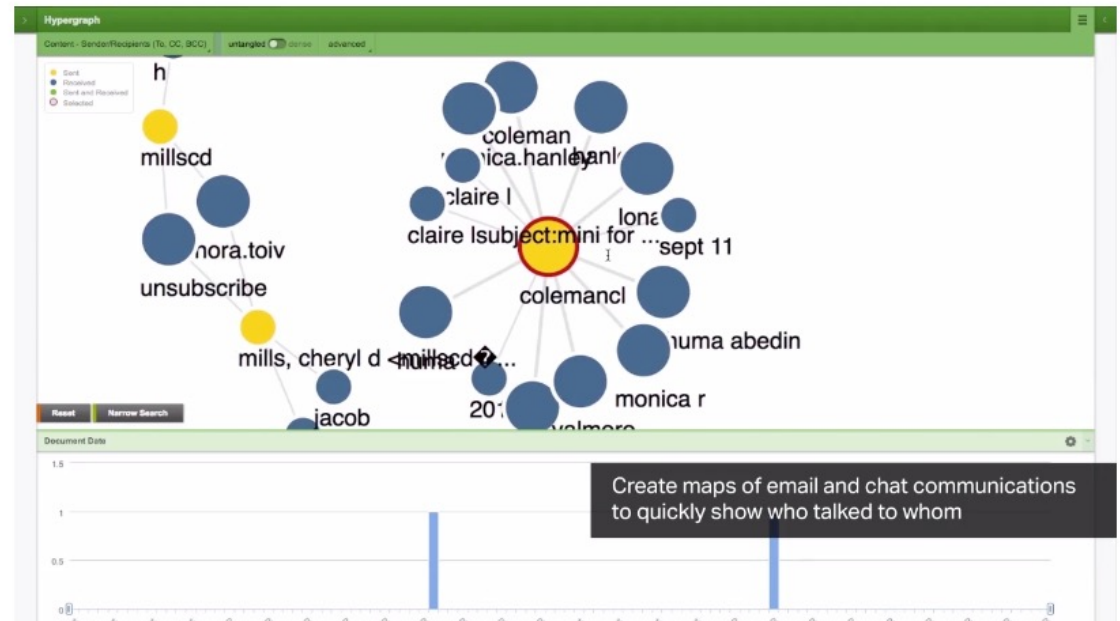
Solutions de Discovery

Avec les logiciels et services eDiscovery et ses fonctions de machine learning, retrouvez plus rapidement les données clés, des données juridico-légales en passant par les données non structurées issues des analyses décisionnelles.

[Demande de démo](#) [Contactez-nous](#)

	bill	office	management	workers	education	art
MSB	bill	management	development	economic	middle	account
MSD	rights	human rights	new york	human	tradition	writing
MSF	coliving	coliving side par	coliving side par	coliving side par	coliving side par	coliving side par
MSH	coliving	coliving side par	coliving side par	coliving side par	coliving side par	coliving side par

Group together similar documents by contextual theme to find relevant content



Solutions de Discovery

Avec les logiciels et services eDiscovery et ses fonctions de machine learning, retrouvez plus rapidement les données clés, des données juridico-légales en passant par les données non structurées issues des analyses décisionnelles.

Contactez-nous



Qu'est qu'un logiciel de Discovery?

Qu'il s'agisse de demandes de renseignements préalables à un litige, d'enquêtes gouvernementales, d'interventions en cas d'atteinte à la protection des données ou d'autres besoins juridiques et de conformité, les logiciels d'eDiscovery aident les entreprises à protéger, recueillir, analyser, classifier, examiner et produire des informations stockées électroniquement (ESI).

Les meilleures solutions logicielles d'eDiscovery exploitent la puissance de l'analyse prédictive, de l'analyse de texte, de la collecte légale de données et des algorithmes de machine learning flexibles. Ces solutions réduisent les volumes de données, automatisent les processus et accélèrent la révision juridique. Les meilleures solutions d'analyse de contrats et moteurs de recherche d'entreprise s'appuient sur ces mêmes technologies fondamentales pour résoudre un éventail encore plus large de problèmes juridiques et commerciaux.

<https://www.opentext.fr/solutions-et-produits/produits/decouverte>

Watson Natural Language Understanding

The natural language processing (NLP) service for advanced text analytics

[Get started free](#)

[→ View demo](#)



Site feedback

What is Watson Natural Language Understanding?

Watson Natural Language Understanding is a cloud native product that uses deep learning to extract metadata from text such as entities, keywords, categories, sentiment, emotion, relations, and syntax.

<h3>Overview</h3>	<h4>Powerful Insight Extraction</h4> <p>Get underneath the topics mentioned in your data by using text analysis to extract keywords, concepts, categories and more.</p> <p>Learn more</p>	<h4>Extensive Language Support</h4> <p>Analyze your unstructured data in more than thirteen languages.</p> <p>Learn more</p>	<h4>High-Accuracy Extraction</h4> <p>Out-of-the-box machine learning models for text mining provide a high degree of accuracy across your content.</p> <p>Learn more</p>
<h3>Why Watson NLU?</h3>	<h4>Deploy Anywhere</h4> <p>Deploy Watson Natural Language Understanding behind your firewall or on any cloud.</p> <p>Learn more</p>	<h4>Domain Customization</h4> <p>Train Watson to understand the language of your business and extract customized insights with Watson Knowledge Studio.</p> <p>Learn more</p>	<h4>Data Control</h4> <p>Maintain ownership of your data with the assurance that your data is safe and secure. IBM will not collect or store your data.</p> <p>Learn more</p>



[Let's talk](#)

SAMPLE INDUSTRY DOMAINS

Legal Financial

Under the **IBM Board Corporate Governance Guidelines**, the **Directors and Corporate Governance Committee** and the **full Board** annually review the financial and other **relationships** between the **independent directors** and IBM as **part of the assessment of director independence**. The **Directors and Corporate Governance Committee** makes **recommendations** to the Board about the independence of non- **management directors**, and the Board determines whether these directors are independent. In **addition to this annual assessment of director independence**, the Board also determines whether these directors are independent.

■ Entities (Out of the total)

Extraction

Entities Keyw

Name
Directors and Corporate Governance Committee
full Board
IBM

SAMPLE INDUSTRY DOMAINS

Legal Financial

Under the **IBM Board Corporate Governance Guidelines**, the **Directors and Corporate Governance Committee** and the **full Board** annually review the financial and other **relationships** between the **independent directors** and IBM as **part of the assessment of director independence**. The **Directors and Corporate Governance Committee** makes **recommendations** to the Board about the independence of non- **management directors**, and the Board determines whether these directors are independent. In **addition to this annual assessment of director independence**, the Board also determines whether these directors are independent.

■ Neutral Entity

Extraction

Sentiment Em

Full Document

Entity Sentiment

Directors and Corporate Governance Committee

full Board

IBM

Keyword Sentiment

IBM Board Corporate Governance Guidelines

independent directors

part of the assessment of director independence

Corporate Governance Committee

Directors

full Board

SAMPLE INDUSTRY DOMAINS | TRY YOUR OWN

Legal Financial Media | Input Text URL

Under the **IBM Board Corporate Governance Guidelines**, the **Directors and Corporate Governance Committee** and the **full Board** annually review the financial and other **relationships** between the **independent directors** and IBM as **part of the assessment of director independence**. The **Directors and Corporate Governance Committee** makes **recommendations** to the Board about the independence of non- **management directors**, and the Board determines whether these directors are independent. In **addition to this annual assessment of director independence**, the Board also determines whether these directors are independent.

■ Sadness ■ Fear ■ Disgust ■ Anger ■ Joy

Extraction Classification Linguistics Custom

Sentiment Emotion Categories

Full Document

Emotion	Percentage
Sadness	1.73%
Joy	32.41%
Fear	3.74%
Disgust	4.37%
Anger	10.53%

Entity Emotion Scores

Directors and Corporate Governance Committee

Emotion	Percentage
Sadness	1.73%
Joy	32.41%
Fear	3.74%



The Word Tree, an Interactive Visual Concordance

Martin Wattenberg and Fernanda B. Viégas

Abstract— We introduce the Word Tree, a new visualization and information-retrieval technique aimed at text documents. A word tree is a graphical version of the traditional "keyword-in-context" method, and enables rapid querying and exploration of bodies of text. In this paper we describe the design of the technique, along with some of the technical issues that arise in its implementation. In addition, we discuss the results of several months of public deployment of word trees on Many Eyes, which provides a window onto the ways in which users obtain value from the visualization.

Index Terms—Text visualization, document visualization, Many Eyes, case study, concordance, information retrieval, search.

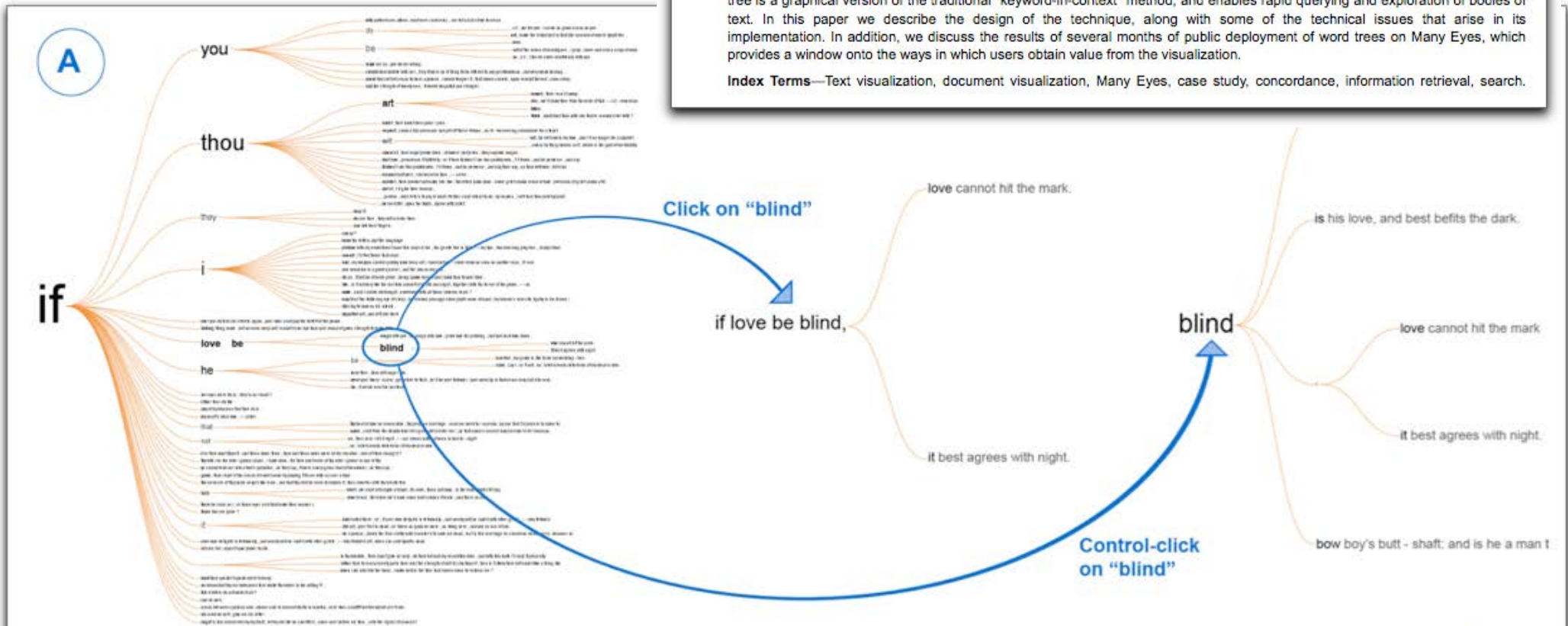
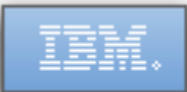
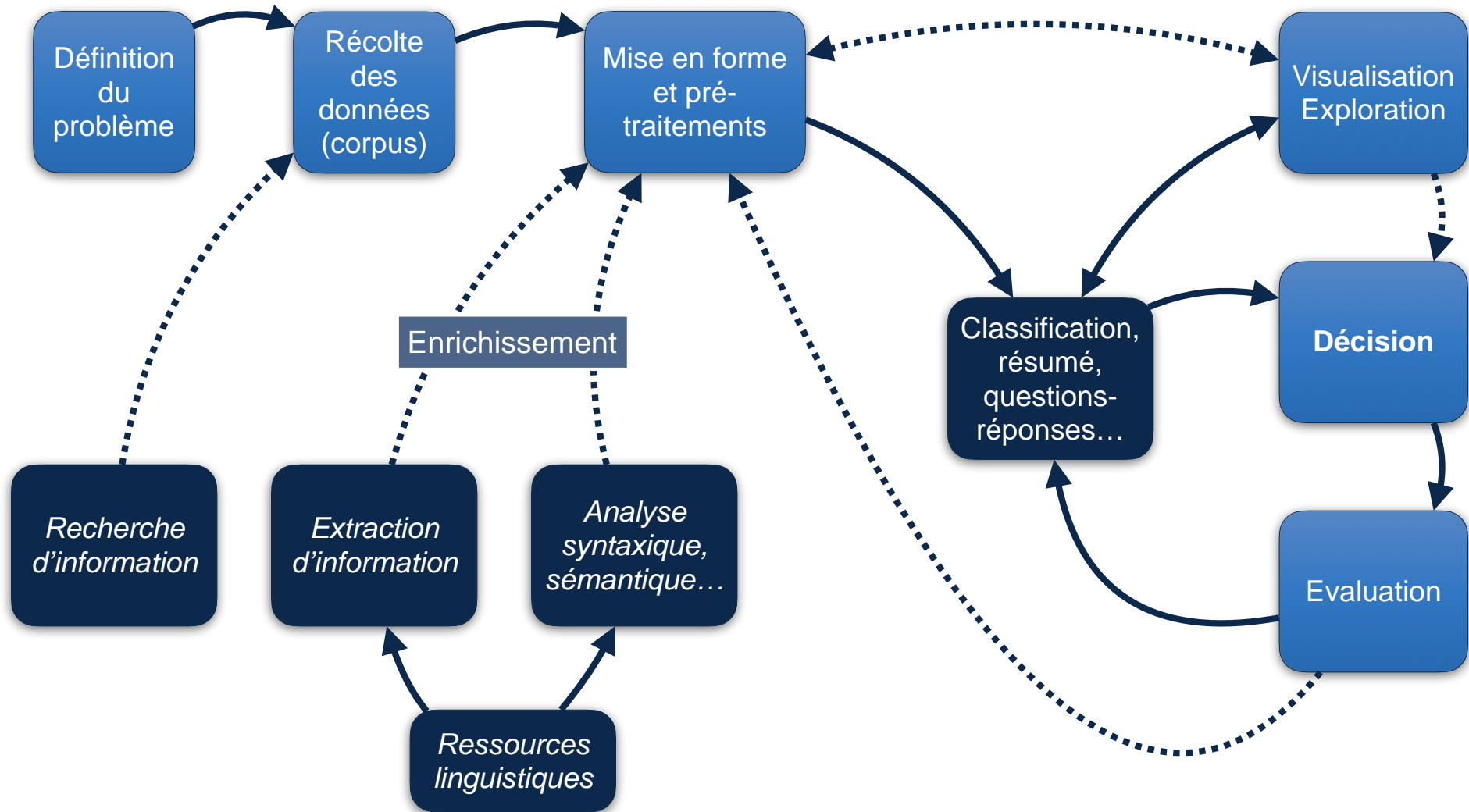


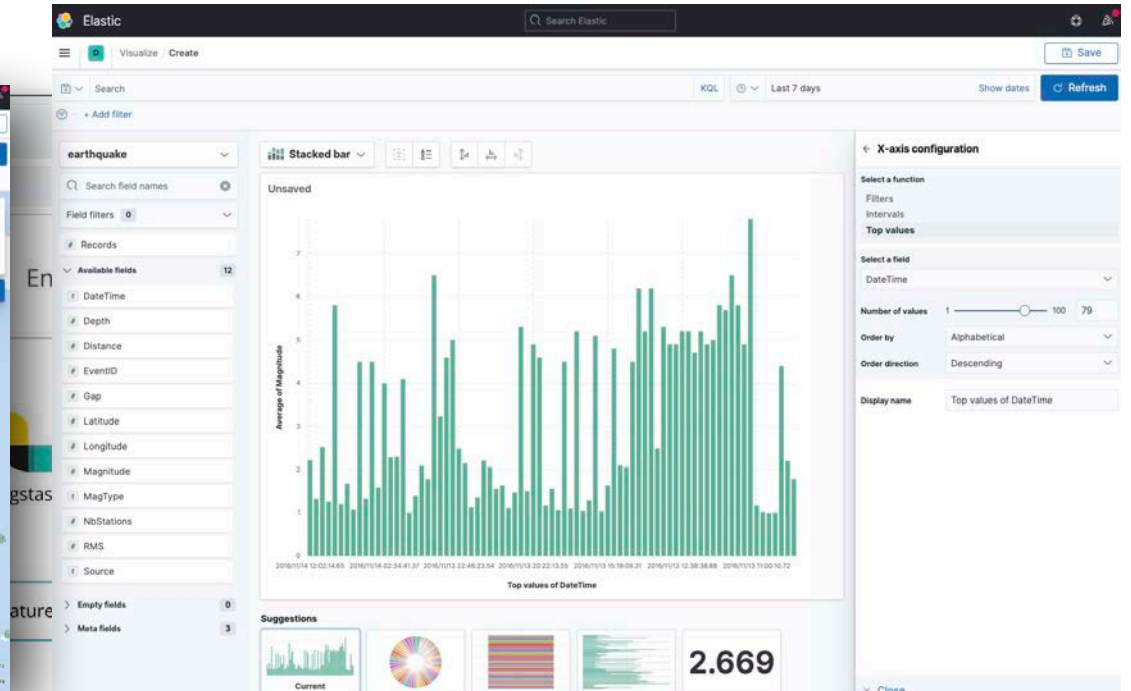
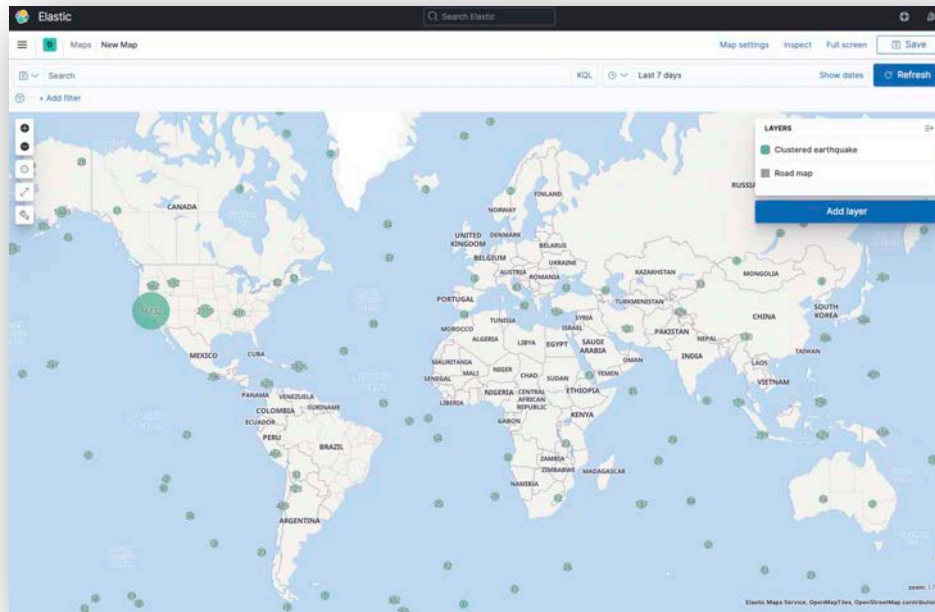
Fig 3. Sequence showing some of the interaction options in the word tree. In figure A, the user has typed the word "if" in *Romeo and Juliet*. In B, the user has clicked on "blind," which appears in one of the branches under "if." This causes the visualization to recenter to the longer phrase "if love be blind." In C, the user Control-clicks on "blind," which causes the visualization to recenter to blind by itself, revealing that there are additional phrases after this term.



Des environnements logiciels pour le développement et l'expérimentation

Un processus de fouille de textes

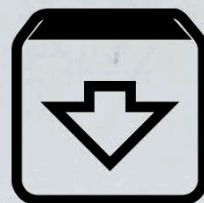




La Suite Elastic

La Suite Elastic, construite sur une fondation open source, vous permet de rechercher, analyser et visualiser, en toute fiabilité et sécurité, ainsi qu'en temps réel, des données issues de n'importe quelle source et sous n'importe quel format.

s'appuie sur Apache Lucene



Téléchargez un corpus ISTEX

Vous êtes membre de l'Enseignement supérieur et de la Recherche et vous souhaitez extraire un corpus de documents ISTEX ?
3 étapes suffisent pour récupérer une archive compressée de votre corpus sur votre disque dur.

1. Requête i

Explicitez le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous :

Équation booléenne i

Identifiants ARK i

Import de fichier i

Exemples i

```
namedEntities.unitex.persName:beethoven AND namedEntities.unitex.placeName:vienna AND namedEntities.unitex.date:"eighteenth century"
```

L'équation saisie correspond à 164 document(s)

Choisir le nombre de documents i : / 164

Choisir les documents classés i :

Par pertinence & qualité Par pertinence Aléatoirement

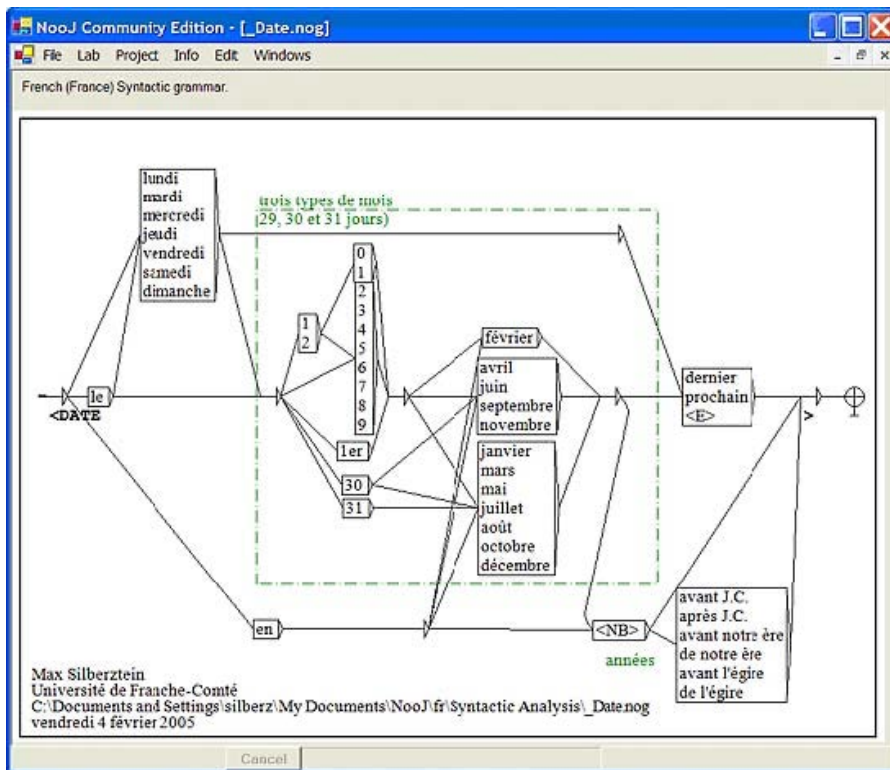
<https://dl.istex.fr>

Traitement de corpus avec grammaires et dictionnaires

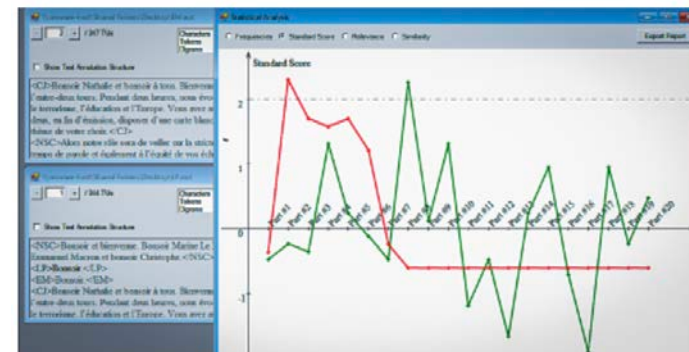
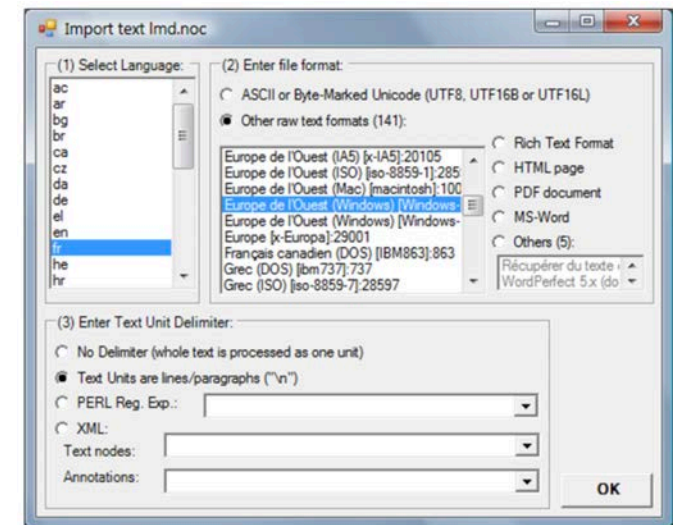


NooJ

A Corpus Processor - A Linguistic Development Environment - A Linguistic Engine for developing Natural Language Processing software Applications.



<https://www.nooj-association.org>



Enrichissement et annotations « linguistiques »

Stanford CoreNLP 4.2.0 (updated 2020-11-16)

<https://corenlp.run>

— Text to annotate —

Parfois dans ces derniers jours d'hiver, nous entrions avant d'aller nous promener dans quelqu'une des petites expositions qui s'ouvraient alors et où Swann, collectionneur de marque, était salué avec une particulière déférence par les marchands de tableaux chez qui elles avaient lieu.

— Annotations —

parts-of-speech x constituency parse x dependency parse x

— Language —

French

Submit

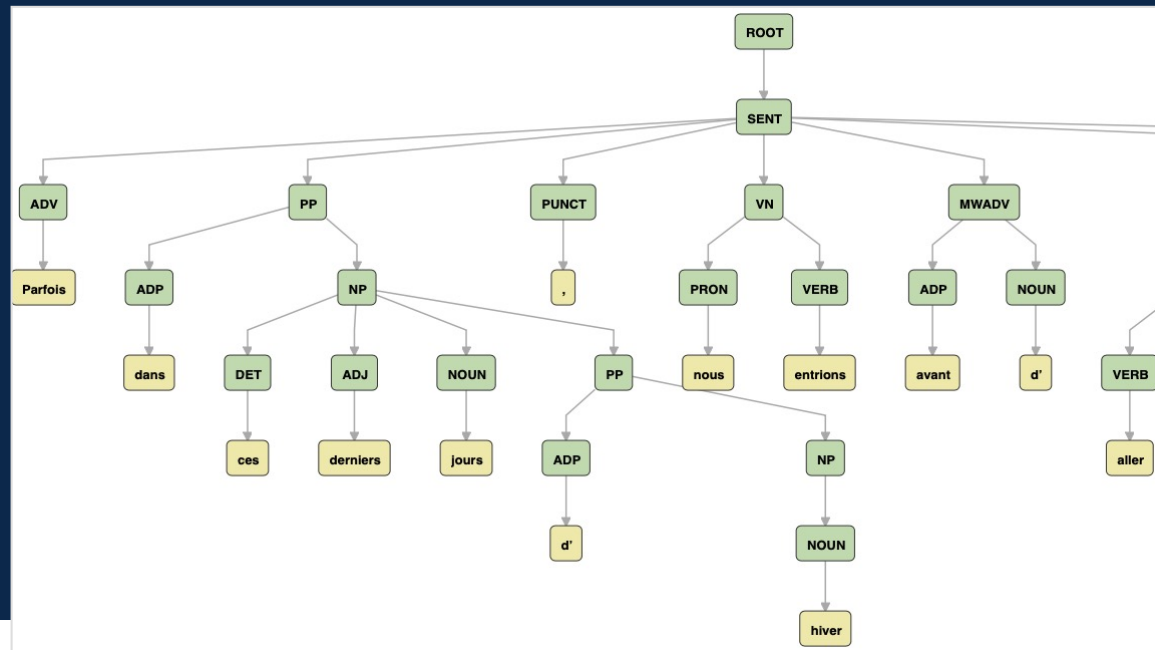
Part-of-Speech:

1 Parfois dans ces derniers jours d' hiver, nous entrions avant d' aller nous promener dans quelqu'une de les petites expositions qui s' ouvraient alors et où Swann, collectionneur de marque, était salué avec une particulière déférence par les marchands de tableaux chez qui elles avaient lieu.

Dépendances

Unités lexicales et parties du discours

Constituants



Structuration automatique de références bibliographiques



cybergeog

Recherche

Index

Auteurs

Mots-clés

Index géographique de référence

Années

Langues

Anniversaire

Les 20 ans de Cybergeog

Rubriques

Bilbo - web

bilbo.openeditionlab.org

```
<?xml version="1.0" encoding="UTF-8"?>
<listBibl>
  <bibl>
    Bortoli M., Cutini V., 1999, Accessibilità urbana e distribuzione delle attività. L'analisi configurazionale del centro storico di Volterra, in Atti della XX Conferenza Italiana di Scienze Regionali, Piacenza, 5-7 Ottobre.
  </bibl>
  <bibl>
    Hill D.M., Bakker J.J., Akers B.L., 1964, An Evaluation of the Needs of the Pedestrian in Downtown, Traffic Research Corporation, Chicago.
  </bibl>
  <bibl>
    Hillier B., 1996, Space is the Machine, Cambridge University Press, Cambridge.
  </bibl>
  <bibl>
    Hillier B., 1999, Why space syntax works, when it looks as though it should not, in Environment & Planning B : Planning and Design, numero speciale monografico sullo Space Syntax Symposium (in corso di pubblicazione).
  </bibl>
  <bibl>
    Hillier B., Hanson J., 1984, The Social Logic of Space, Cambridge University Press, Cambridge.
  </bibl>
  <bibl>
    Hillier B., Penn A., Hanson J., Grajevski, Xu J., 1993, Natural Movement : or, Configuration and Attraction in Urban Pedestrian Movement, in Enviroment & Planning B, Planning and Design, vol. 20.
  </bibl>
  <bibl>
    Hoel L.A., 1968, Pedestrian Travel Rates in Central Business Districts, in Traffic Engineering and Control, January, 10-13.
  </bibl>
  <bibl>
    Lautso K., Murola P., 1974, A Study of Pedestrian Traffic in Helsinki, in Traffic Engineering and Control, January, 446-449.
  </bibl>
  <bibl>
    O'Flaherty C.A., Parkinson M.H., 1972, Movement on a City Centre Footway, in Traffic Engineering and Control, February, 434-438.
  </bibl>
  <bibl>
    Pushkarev B., Zupan J., 1975, Urban Space for Pedestrians, MIT Press, Cambridge, MA.
  </bibl>
</listBibl>
```

DOI :

Corpus 1 (bibliography) : / Corpus 2 (notes) :

Annotate - Reset

Test corpus 1 Test corpus 2

Type:

Bortoli M., Cutini V., 1999, Accessibilità urbana e distribuzione delle attività. L'analisi configurazionale del centro storico di Volterra, in Atti della XX Conferenza Italiana di Scienze Regionali, Piacenza, 5-7 Ottobre.

Hill D.M., Bakker J.J., Akers B.L., 1964, An Evaluation of the Needs of the Pedestrian in Downtown, Traffic Research Corporation, Chicago.

Hillier B., 1996, Space is the Machine, Cambridge University Press, Cambridge.

Hillier B., 1999, Why space syntax works, when it looks as though it should not, in Environment & Planning B : Planning and Design, numero speciale monografico sullo Space Syntax Symposium (in corso di pubblicazione).

Hillier B., Hanson J., 1984, The Social Logic of Space, Cambridge University Press, Cambridge.

Hillier B., Penn A., Hanson J., Grajevski, Xu J., 1993, Natural Movement : or, Configuration and Attraction in Urban Pedestrian Movement, in Enviroment & Planning B, Planning and Design, vol. 20.

Hoel L.A., 1968, Pedestrian Travel Rates in Central Business Districts, in Traffic Engineering and Control, January, 10-13.

Lautso K., Murola P., 1974, A Study of Pedestrian Traffic in Helsinki, in Traffic Engineering and Control, January, 446-449.

O'Flaherty C.A., Parkinson M.H., 1972, Movement on a City Centre Footway, in Traffic Engineering and Control, February, 434-438.

Pushkarev B., Zupan J., 1975, Urban Space for Pedestrians, MIT Press, Cambridge, MA.

spatial simulation », *Computers, Environment and Urban Systems*, vol. 32, No.6, 417-430.
DOI : 10.1016/j.compenvurbsys.2008.09.004

Test : <http://bilbo.openeditionlab.org>

Sources : <http://github.com/OpenEdition/bilbo>



AlvisAE (1)

Annotation manuelle collective de textes.

The screenshot displays the AlvisAE web interface for manual text annotation. The top bar shows the user 'claire' and the project 'annotation : [train+dev] BTID-30042'. The left sidebar contains a taxonomic tree with categories like 'ferret', 'swine', 'herbivore', 'ruminant', 'bovine', 'small ruminant', 'ruminant livestock (2)', 'dog (2)', 'domestic animal', 'cat', 'farm animal', 'sheep (2)', and 'cattle (2)'. The main text area shows a paragraph about 'Mycoplasma agalactiae' with terms highlighted in colored boxes: 'sheep', 'goats', 'Mycoplasmas', 'bacterium Mycoplasma agalactiae', 'small ruminants', 'mycoplasmas', 'Mollicutes', 'bacteria', 'Gram-positive bacteria with low G+C content', 'Firmicutes', 'Clostridia', and 'Bacilli'. The bottom section, titled 'Annotations', contains a table with columns for 'Id', 'Annotation Set', 'K', 'Type', 'Details', and 'Vis'.

Id	Annotation Set	K	Type	Details	Vis
9cb98.	[imported] Annotation		Localization	Bacterium (Bacteria Mycoplasma agalactiae) + Localization (Geographical Europe)	
78218.	[imported] Annotation		Localization	Bacterium (Bacteria M. mycoides subsp. mycoides SC) + Localization (Host small ruminants)	
40bbc.	[imported] Annotation		Localization	Bacterium (Bacteria Mycoplasma agalactiae) + Localization (Host sheep)	
0923d.	[imported] Annotation		Localization	Bacterium (Bacteria Mycoplasma agalactiae) + Localization (Host small ruminants)	

prodigy

<https://prodi.gy/>

Radically efficient machine tea
An annotation tool powered
by active learning.

Named Entity Recognition

RTL CJK character-based

Try it live and highlight entities!

LOCATION 1 EVENT 2 DATE 3

وفي الفترة 1944-1945 افلح الجيش الكندي الأول في تحرير معظم أراضي هولندا LOCATION ، وكان يضم في صفوفه قوات كندية وبريطانية ويولندية . لكن سرعان ما بات لزاماً على الهولنديين بنهاية الحرب الأوروبية أن يحاربوا مقاتلي الثورة الوطنية الإندونيسية

SOURCE: ar.wikipedia.org/wiki/%D9%87%D9%88%D9%84%D9%86%D8%AF%D8%A7

Label any text, in any language or script

Prodigy lets you use token boundaries for faster and more consistent annotation, but it's also fully flexible: you can annotate from the character up if your task requires it. No matter what language or writing system you're working with, if it's text, Prodigy can help you annotate it.

Bootstrap with powerful patterns

Prodigy is a fully scriptable annotation tool, letting you **automate as much as possible** with custom rule-based logic. You don't want to waste time labeling every instance of common entities like "New York" or "the United States" by hand. Instead, give Prodigy rules or a list of examples, review the entities in context and annotate the exceptions. As you annotate, a statistical model can learn to suggest similar entities, generalising beyond your initial patterns.

patterns.json

```
{ "pattern": [ { "lower": "new" }, { "lower": "york" } ], "label": "CITY" }  
{ "pattern": [ { "lower": "berlin" } ], "label": "CITY" }
```

I live in New York CITY .



Classification



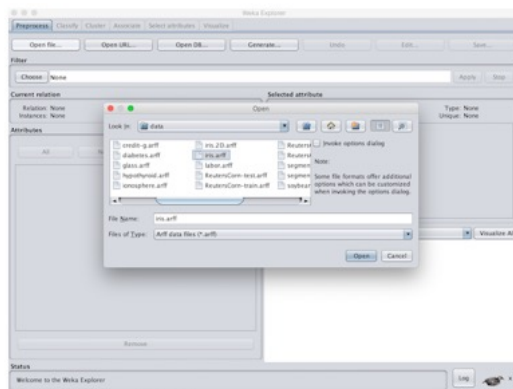
WEKA

The workbench for machine learning

Weka is tried and tested open source machine learning : that can be accessed through a graphical user interface, terminal applications, or a Java API. It is widely used for t

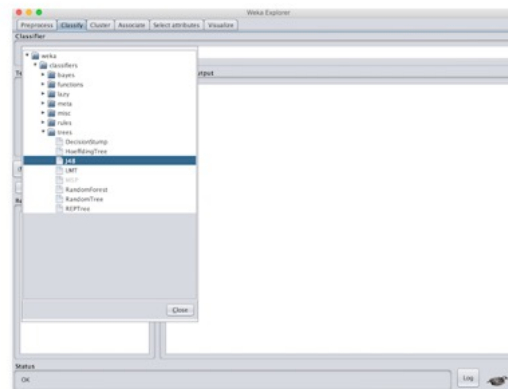
Machine Learning without Programming

Weka can be used to build machine learning pipelines, train classifiers, and run evaluations without having to write a single line of code:



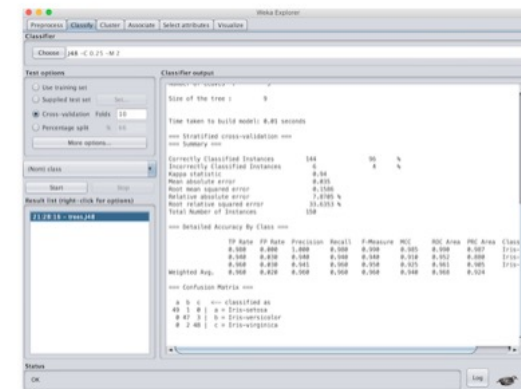
Open a dataset

First, we open the [dataset](#) that we would like to evaluate.



Choose a classifier

Second, we select a learning algorithm to use, e.g., the J48 classifier, which learns decision trees.



Evaluate predictive accuracy

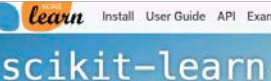
Finally, we run a 10-fold cross-validation evaluation and obtain an estimate of predictive performance.

Des bibliothèques Python (ou Java, C++, Swift)



Welcome to Apache OpenNLP

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.



scikit-learn

Machine Learning in Python

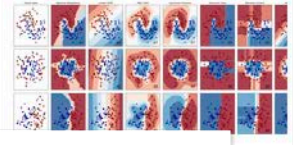
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition, fraud detection, and more...

Algorithms: SVM, nearest neighbors, random forest, and more...

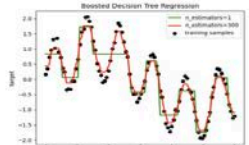


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices, and more...

Algorithms: SVR, nearest neighbors, random forest, and more...




Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes, and more...

Algorithms: k-Means, spectral clustering, mean-shift, and more...




CoreNLP



TensorFlow

Google is committed to

Une plate-forme Open Source de bout en bout dédiée au machine learning



Keras

Simple. Flexible. Powerful.

[Get started](#) [Guides](#) [API docs](#)

```
from tensorflow import keras
from tensorflow.keras import layers

# Instantiate a trained vision model
vision_model = keras.applications.ResNet50()


# Use it to preprocess images using the trained vision model
video_input = keras.Input(shape=(180, 180, 3))
encoded_frame_sequence = layers.TimeDistributed(vision_model)(video_input)
encoded_video = layers.LSTM(256)(encoded_frame_sequence)

# Use it to preprocess questions using the trained question model
question_input = keras.Input(shape=(180, 1), dtype='int32')
embedded_question = layers.Embedding(10000, 256)(question_input)
encoded_question = layers.LSTM(256)(embedded_question)

# Use this to create a question answering model
merged = keras.layers.concatenate([encoded_video, encoded_question])
output = keras.layers.Dense(1000, activation='softmax')(merged)
video_qa_model = keras.Model([video_input, question_input], output)
```

Deep learning for humans.

Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear & actionable error messages. It also has extensive documentation and developer guides.




PyTorch

Get Started Ecosystem Mobile Blog

FROM RESEARCH TO PRODUCTION

An open source machine learning framework that accelerates the path from research prototyping to production deployment.



AllenNLP

A natural language processing platform for building state-of-the-art models.

Answer a question ^

Reading Comprehension

Visual Question Answering

Annotate a sentence ^

Named Entity Recognition

Open Information Extraction

Sentiment Analysis

Dependency Parsing

Constituency Parsing

Semantic Role Labeling

Annotate a passage ^

Coreference Resolution

Generate a passage ^

Language Modeling

Masked Language Modeling

Compare two sentences ^

Textual Entailment

Named Entity Recognition

Named Entity Recognition is the task of identifying named entities (people, locations, organizations, etc.) in the input text.

Model

Fine Grained Named Entity Recognition

This model identifies a broad range of 16 semantic types in the input text. It is a reimplementaion of Lample (2016) and uses a biLSTM with a CRF layer, character embeddings and ELMo embeddings.

[TaskDemo](#)

[Model Card](#)

[Model Usage](#)

Example Inputs

When I told John that I wanted to move to Alaska, he warned me that I'd have trouble finding a Starbucks there.

Sentence

When I told John that I wanted to move to Alaska, he warned me that I'd have trouble finding a Starbucks there.

Run Model

Model Output

Share

Entities

When I told John that I wanted to move to Alaska, he warned me that I'd have trouble finding a Starbucks there .

PERSON GPE ORG

[Reading Comprehension](#)[Visual Question Answering](#)[✍ Annotate a sentence](#) ^[Named Entity Recognition](#)[Open Information Extraction](#)[Sentiment Analysis](#)[Dependency Parsing](#)[Constituency Parsing](#)[Semantic Role Labeling](#)[☰ Annotate a passage](#) ^[Coreference Resolution](#)[🧩 Generate a passage](#) ^[Language Modeling](#)[Masked Language Modeling](#)[🔗 Compare two sentences](#) ^[Textual Entailment](#)

Language Modeling

Language modeling is the task of determining the probability of a given sequence of words occurring in a sentence.

Model

GPT2-based Next Token Language Model

This is the public 345M parameter OpenAI GPT-2 language model for generating sentences. The model embeds some input tokens, contextualizes them, then predicts the next word, computing a loss against known target. If `BeamSearch` is given, this model will predict a sequence of next tokens.

[TaskDemo](#)[Model Card](#)

Example Inputs

Sentence

[Run Model](#)

Model Output

[Share](#)

Prediction	Score
The doctor ran to the emergency room to see the patient. ↵" ...	 99,1 %
The doctor ran to the emergency room to see the girl. She was crying ...	 0,6 %
The doctor ran to the emergency room to see the injured victim. ↵	 0,2 %

Presets

-  Explain a word
-  De-Jargonizer
-  Predict the outcome
-  Generate code
-  Classify news topics
-  Summarize restaurant reviews
-  Blog post ideation
-  Sports trivia
-  Convert text to table
-  Table question answering
-  Product description generator
-  Python to Javascript
-  Openness classifier
-  Generate catchy headlines

Canvas [Quickstart](#)

 Clear all  Share

Predict common sense results of the following actions.

--

Action: I didn't water the plant for 3 weeks.
Result: The plant died.

--

Action: I went to school.
Result: I got a diploma.

--

Action: I left the AC on all day.
Result: I got a high utility bill.

--

Action: I helped my neighbors when their car broke down.
Result: My neighbors were grateful.

--

Action: I put the ice cream outside for an hour.
Result: **The ice cream melted.**

--|


 Generate



95 / 2048 

Configuration

Model

j1-jumbo (178B) 

Max completion length 20



Temperature 0.8



Top P 0.98



Stop sequences

Alternative tokens

Chat Open ended conversation with an AI assist...

Grammar correction Corrects sentences into standard English.

Natural language to OpenAI API Create code to call to the OpenAI API usin...

English to French Translates English text into French.

SQL translate Translate natural language to SQL queries.

Classification Classify items into categories via example.

Movie to Emoji Convert movie titles into emoji.

Translate programming languages Translate from one programming language ...

Explain code Explain a complicated piece of code.

Factual answering Guide the model towards factual answering ...

Q&A Answer questions based on existing know

Summarize for a 2nd grader Translates difficult text into simpler conce

Text to command Translate text into programmatic commar

Natural language to Stripe API Create code to call the Stripe API using n

Parse unstructured data Create tables from long form text

Python to natural language Explain a piece of Python code in human

Calculate Time Complexity Find the time complexity of a function.

Advanced tweet classifier Advanced sentiment detection for a piece

Keywords Extract keywords from a block of text.

Ad from product description Turn a product description into ad copy.

Product name generator Create product names from examples word...

Python bug fixer Find and fix bugs in source code.

JavaScript helper chatbot Message-style bot that answers JavaScript ...

Science fiction book list maker Create a list of items for a given topic.

Airport code extractor Extract airport codes from text.

Extract contact information Extract contact information from a block of ...

Friend chat Emulate a text message conversation.

Write a Python docstring An example of how to create a docstring for ...

JavaScript one line function Turn a JavaScript function into a one liner.

Third-person converter Converts first-person POV to the third-pers...

VR fitness idea generator Create ideas for fitness and virtual reality g...

Essay outline Generate an outline for a research topic.

TL;DR summarization Summarize text by adding a 'tl;dr:' to the en...

Spreadsheet generator Create spreadsheets of various kinds of dat...

ML/AI language model tutor Bot that answers questions about language...

Tweet classifier Basic sentiment detection for a piece of text.

SQL request Create simple SQL queries.

JavaScript to Python Convert simple JavaScript expressions into ...

Mood to color Turn a text description into a color.

Analogy maker Create analogies. Modified from a communi...

Micro horror story creator Creates two to three sentence short horror ...

Notes to summary Turn meeting notes into a summary.

ESRB rating Categorize text based upon ESRB ratings.

Recipe generator Create a recipe from a list of ingredients.



Build next-gen apps with OpenAI's powerful models.

OpenAI's API provides access to GPT-3, which performs a wide variety of natural language tasks, and Codex, which translates natural language to code.

Pricing

Simple and flexible. Only pay for what you use.

JOIN THE WAITLIST

Per-model prices

Ada <small>Fastest</small>	Babbage	Curie	Davinci <small>Most powerful</small>
\$0.0008 /1K tokens	\$0.0012 /1K tokens	\$0.0060 /1K tokens	\$0.0600 /1K tokens

Multiple models, each with different capabilities and price points. **Ada** is the fastest model, while **Davinci** is the most powerful.

Prices are per 1,000 tokens. You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens.

NLTK Corpora

NLTK has built-in support for dozens of corpora and trained models, as listed below. To use these within NLTK we recommend that you please consult the README file included with each corpus for further information.

NLTK 3.5 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

- 1. [Unicode Version 7.0.0 character properties in Perl](#) [[download](#) | [source](#)]
id: unicode; size: 100266; author: ; copyright: ; license: ;
- 2. [aligner \(Sultan et al. 2015\) subset of the Paraphrase Database](#). [[download](#) | [source](#)]
id: aligner; size: 711; author: ; copyright: ; license: Creative Commons Attribution 3.0 Unported (CC-BY);
- 3. [\[download | source\]](#)
id: 3; author: Jan Strunk; copyright: ; license: ;
- 4. [Corpus de Sufixos da Língua Portuguesa](#) [[download](#) | [source](#)]
id: portuguese; author: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christian Huyck; copyright: ; license: ;
- 5. [es](#) [[download](#) | [source](#)]
id: es; size: 100510; author: ; copyright: ; license: ;
- 6. [\[download | source\]](#)
id: 6; size: 6785405; author: ; copyright: ; license: ;
- 7. [\[download | source\]](#)
id: 7; size: 13404747; author: ; copyright: ; license: ;
- 8. [\[download | source\]](#)
id: 8; size: 10961490; author: ; copyright: ; license: ;
- 9. [\[download | source\]](#)
id: 9; size: 24516205; author: ; copyright: ; license: ;
- 10. [\[download | source\]](#)
id: 10; size: 49396025; author: ; copyright: ; license: ;
- 11. [Evaluation data from WMT15](#) [[download](#) | [source](#)]
id: wmt15_eval; size: 383096; author: ; copyright: ; license: ;
- 12. [Grammars for Spanish](#) [[download](#) | [source](#)]
id: spanish_grammars; size: 4047; author: Kepa Sarasola; copyright: ; license: ;
- 13. [Sample Grammars](#) [[download](#) | [source](#)]
id: sample_grammars; size: 20293; author: ; copyright: ; license: ;
- 14. [Large context-free and feature-based grammars for parser comparison](#) [[download](#) | [source](#)]
id: large_grammars; size: 283747; author: ; copyright: ; license: See the individual grammar files;
- 15. [Grammars from NLTK Book](#) [[download](#) | [source](#)]
id: book_grammars; size: 9103; author: Ewan Klein; copyright: ; license: ;
- 16. [Grammars for Basque](#) [[download](#) | [source](#)]
id: basque_grammars; size: 4704; author: Kepa Sarasola; copyright: ; license: ;

Some simple things you can do with NLTK

Tokenize and tag some text:

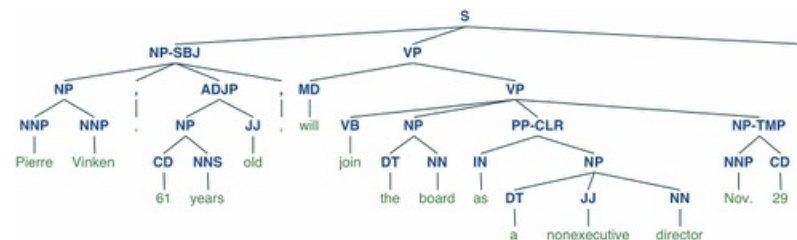
```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

Identify named entities:

```
>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [(('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'),
('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN')),
Tree('PERSON', [(('Arthur', 'NNP')]),
('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'),
('very', 'RB'), ('good', 'JJ'), (',', ','))])
```

Display a parse tree:

```
>>> from nltk.corpus import treebank
>>> t = treebank.parsed_sents('wsj_0001.mrg')[0]
>>> t.draw()
```



```
In [2]: from nltk.twitter import Twitter
tw = Twitter()
tw.tweets(keywords='love, hate', limit=10) #sample from the public stream
```

Sana magkakaisa na ang mga Kapamilya at Kapuso. Spread love, not hate
 #ShowtimeKapamiIyaDay #ALDubEBforLOVE
 @Real_Liam_Payne Please follow me , you mean the world to me and words can't describe how much i love you x3186
 Love my ugly wife
 RT @ansaberano: We Found Love
 #PushAwardsLizQuen
 RT @yungunmei: people want to fall in love but don't understand the concept
 I don't care, I love It #EMABiggestFans1D
 RT @bryan white: I'm not in the Philippines Yet but we are making a very BIG announcement in 2 days! Get ready! Love
 you! #GGMY #ALDubEBfor...
 I whole heartedly HATE @lakiamichelle like really HATE her 😞 who wants to be her friend because I DONT
 RT @lahrose23: I love yu to https://t.co/dfsRwSp1IC
 RT @alone_in_woods: ahoj, já jsem tvůj pes a tohle je náš love song /// Zrní - Já jsem tvůj pes https://t.co/7L0XPHeA
 2d via @YouTube
 Written 10 Tweets

Sentiment Analysis

```
>>> from nltk.classify import NaiveBayesClassifier
>>> from nltk.corpus import subjectivity
>>> from nltk.sentiment import SentimentAnalyzer
>>> from nltk.sentiment.util import *

>>> n_instances = 100
>>> subj_docs = [(sent, 'subj') for sent in subjectivity.sents(categories='subj')[n_instances]]
>>> obj_docs = [(sent, 'obj') for sent in subjectivity.sents(categories='obj')[n_instances]]
>>> len(subj_docs), len(obj_docs)
(100, 100)
```

Each document is represented by a tuple (sentence, label). The sentence is tokenized, so it is represented by a list of strings:

```
>>> subj_docs[0]
(['smart', 'and', 'alert', ',', 'thirteen', 'conversations', 'about', 'one',
 'thing', 'is', 'a', 'small', 'gem', '.'], 'subj')
```

We separately split subjective and objective instances to keep a balanced uniform class distribution in both train and test sets.

```
>>> train_subj_docs = subj_docs[:80]
>>> test_subj_docs = subj_docs[80:100]
>>> train_obj_docs = obj_docs[:80]
>>> test_obj_docs = obj_docs[80:100]
>>> training_docs = train_subj_docs+train_obj_docs
>>> testing_docs = test_subj_docs+test_obj_docs

>>> sentim_analyzer = SentimentAnalyzer()
>>> all_words_neg = sentim_analyzer.all_words([mark_negation(doc) for doc in training_docs])
```

We use simple unigram word features, handling negation:

```
>>> unigram_feats = sentim_analyzer.unigram_word_feats(all_words_neg, min_freq=4)
>>> len(unigram_feats)
83
>>> sentim_analyzer.add_feat_extractor(extract_unigram_feats, unigrams=unigram_feats)
```

We apply features to obtain a feature-value representation of our datasets:

```
>>> training_set = sentim_analyzer.apply_features(training_docs)
>>> test_set = sentim_analyzer.apply_features(testing_docs)
```

We can now train our classifier on the training set, and subsequently output the evaluation results:

```
>>> trainer = NaiveBayesClassifier.train
>>> classifier = sentim_analyzer.train(trainer, training_set)
Training classifier
>>> for key,value in sorted(sentim_analyzer.evaluate(test_set).items()):
...     print('{0}: {1}'.format(key, value))
Evaluating NaiveBayesClassifier results...
Accuracy: 0.8
F-measure [obj]: 0.8
F-measure [subj]: 0.8
Precision [obj]: 0.8
Precision [subj]: 0.8
Recall [obj]: 0.8
Recall [subj]: 0.8
```

<http://www.nltk.org>

Industrial-Strength Natural Language Processing

IN PYTHON

Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

GET STARTED

Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

FACTS & FIGURES

Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

READ MORE


```
Edit the code & try spaCy spaCy v3.0 · Python 3 · via Binder

# pip install -U spacy
# python -m spacy download en_core_web_sm
import spacy

# Load English tokenizer, tagger, parser and NER
nlp = spacy.load("en_core_web_sm")

# Process whole documents
text = ("When Sebastian Thrun started working on self-driving cars at "
        "Google in 2007, few people outside of the company took him "
        "seriously. "I can tell you very senior CEOs of major American "
        "car companies would shake my hand and turn away because I wasn't "
        "worth talking to," said Thrun, in an interview with Recode earlier "
        "this week.")
doc = nlp(text)

# Analyze syntax
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])

# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)

RUN
```

```
Noun phrases: ['Sebastian Thrun', 'self-driving cars', 'Google', 'few people', 'th
e company', 'him', 'I', 'you', 'very senior CEOs', 'major American car companies',
'my hand', 'I', 'Thrun', 'an interview', 'Recode']
Verbs: ['start', 'work', 'drive', 'take', 'tell', 'shake', 'turn', 'be', 'talk', '
say']
Sebastian Thrun PERSON
2007 DATE
American NORP
Thrun PERSON
Recode PERSON
earlier this week DATE
```

Features

- ✓ Support for **69+ languages**
- ✓ **58 trained pipelines** for 18 languages
- ✓ Multi-task learning with pretrained **transformers** like BERT
- ✓ Pretrained **word vectors**
- ✓ State-of-the-art speed
- ✓ Production-ready **training system**
- ✓ Linguistically-motivated **tokenization**
- ✓ Components for **named entity** recognition, part-of-speech tagging, dependency parsing, sentence segmentation, **text classification**, lemmatization, morphological analysis, entity linking and more
- ✓ Easily extensible with **custom components** and attributes
- ✓ Support for custom models in **PyTorch**, **TensorFlow** and other frameworks
- ✓ Built in **visualizers** for syntax and NER
- ✓ Easy **model packaging**, deployment and workflow management
- ✓ Robust, rigorously evaluated accuracy

Features

In the documentation, you'll come across mentions of spaCy's features and capabilities. Some of them refer to linguistic concepts, while others are related to more general machine learning functionality.

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) Tagging	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat".
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

CONCLUSION : LE TDM...

Concerne et impacte :

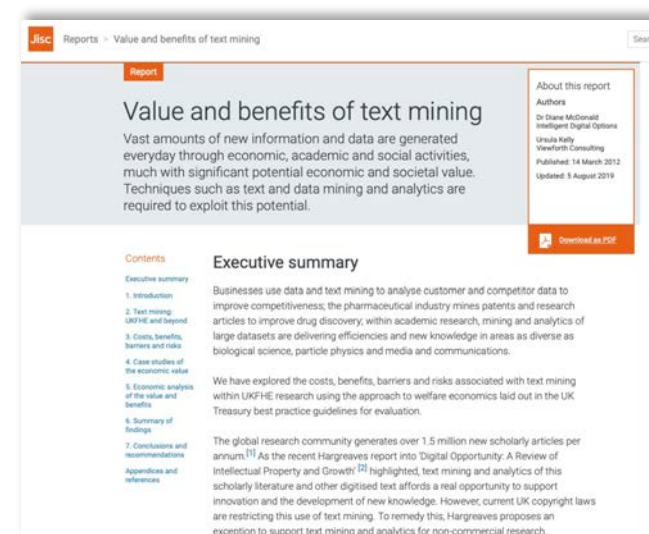
- La recherche scientifique dans son ensemble
- La société au travers d'applications du quotidien

Nécessite :

- un corpus cible, des ressources de spécialité
- d'intégrer différents composants logiciels, modèles, APIs
- un scénario et une référence pour apprendre et évaluer

Faisable si :

- les composants sont interopérables, les métadonnées compatibles
- l'intégration de différents composants logiciels est possible
(ou s'il existe déjà une brique logicielle répondant au besoin)



<https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>

De nombreux ateliers et compétitions



The CLEF Initiative
Conference and Labs of the Evaluation Forum

<http://www.clef-initiative.eu/>

1. **ARQMath: Answer Retrieval for Questions on Math**
2. **BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering**
3. **CheckThat!: Automatic Identification and Verification of Claims**
4. **ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents**
5. **eHealth: Retrieving and Making Sense of Medical Content**
6. **eRisk: Early Risk Prediction on the Internet**
7. **HIPE: Identifying Historical People, Places and other Entities**
8. **ImageCLEF: Multimedia Retrieval in Medicine, Lifelogging, and Internet**
9. **LifeCLEF: Multimedia Retrieval in Nature**
10. **LiLAS: Living Labs for Academic Search**
11. **PAN: Stylometry and Digital Text Forensics**
12. **Touché: Argument Retrieval**

LifeCLEF - Biodiversity identification and prediction

ProtestNews - Extracting Protests from News

eHealth

CENTRE@CLEF

eRISK - Early Risk prediction on the Internet

PAN Lab on Digital Text Forensics and Stylometry

CheckThat! - Automatic Identification and Verification of Claims

PIR-CLEF - Evaluation of personalised IR

SemEval-2021

The 15th International Workshop on Semantic Evaluation

<https://semeval.github.io/SemEval2021/tasks.html>

Lexical semantics

- **Task 1: Lexical Complexity Prediction** ([email organizers] [mailing list])
Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, Marcos Zampieri
- **Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation** ([email organizers])
NOTE: new competition website!
Federico Martelli, Najla Kalach, Gabriele Tola, Roberto Navigli
- **Task 3: Span- and Dependency-based Multilingual and Cross-lingual Semantic Role Labeling**
- **Task 4: Reading Comprehension of Abstract Meaning** ([email organizers] [mailing list])
Boyuan Zheng, Xiaoyu Yang, Yu-Ping Ruan, Quan Liu, Zhen-Hua Ling, Si Wei, Xiaodan Zhu

Social factors & opinion

- **Task 5: Toxic Spans Detection** ([email organizers] [mailing list])
John Pavlopoulos, Ion Androutsopoulos, Jeffrey Sorensen, Léo Laugier
- **Task 6: Detection of Persuasive Techniques in Texts and Images** *Updated website* ([email organizers] [mailing list])
Giovanni Da San Martino, Hamed Firooz, Preslav Nakov, Fabrizio Silvestri
- **Task 7: HaHackathon: Detecting and Rating Humor and Offense** ([email organizers])
NOTE: new competition website!
J. A. Meaney, Steven Wilson, Walid Magdy, Luis Chiruzzo

Information in scientific & clinical text

- **Task 8: MeasEval: Counts and Measurements** ([email organizers] [mailing list])
Corey Harper, Jessica Cox, Ron Daniel, Paul Groth, Curt Kohler, Antony Scerri
- **Task 9: Statement Verification and Evidence Finding with Tables** ([email organizers] [mailing list])
Nancy Xin Ru Wang, Sara Rosenthal, Marina Danilevsky, Diwakar Mahajan
- **Task 10: Source-Free Domain Adaptation for Semantic Processing** ([email organizers] [mailing list])
Steven Bethard, Egoitz Laparra, Timothy Miller, Özlem Uzuner
- **Task 11: NLPContributionGraph** ([email organizers] [mailing list])
Jennifer D'Souza, Sören Auer, Ted Pedersen

De nombreux ateliers et compétitions (2)

The screenshot shows the Kaggle interface for the 'Getting Started Prediction Competition' titled 'Natural Language Processing with Disaster Tweets'. The main heading is 'Predict which Tweets are about real disasters and which ones are not'. It indicates that the competition is ongoing with 1,325 teams. The navigation menu includes Overview, Data, Code, Discussion, Leaderboard, Datasets, and Rules. The description section is partially visible, starting with 'Welcome to one of our "Getting Started" competit'.

<https://www.kaggle.com/c/nlp-getting-started>

The screenshot shows the CodaLab Competitions website at <https://competitions.codalab.org/competitions/>. It features a search bar and a list of various competitions:

- Evaluating grammatical error corrections**: Organized by cnapoles. This "competition" contains different evaluation metrics commonly used for GEC and allows users to score their systems with these metrics ...
- ADoBo — Automatic Detection of Borrowings**: Organized by lea. Detecting emerging borrowings from English into Spanish (words like 'smartphone' or 'fake news') that appear in the Spanish press
- Interspeech Shared Task: Automatic Speech Recognition for Non-Native Children's Speech**: Organized by cleong. A joint shared task between FBK, ETS and Cambridge
- ICDAR 2021 Competition on Scene Video Text Spotting**: Organized by Embers. To support this competition, we extend the Larger-Scale Video Text Dataset released in YORO [1], and release a dataset containing ...
- EmoEvalEs@IberLEF 2021**: Organized by amontejo. Workshop en Emotion detection and Evaluation for Spanish



Thirteenth Text Analysis Conference (TAC 2020)

Evaluation: August 2020 - January 2021
Workshop: February 22-23, 2021

- Epidemic Question Answering (EPIC-QA)**
 The goal of the EPIC-QA track is to evaluate systems on their ability to answer questions about COVID-19, related coronaviruses, and the recommended response to the pandemic. The challenge is to return expert-level answers as expected by the scientific community.
Track coordinators: Dina Demner-Fushman (ddemner@mail.nih.gov)
Home page: https://bionlp.nlm.nih.gov/epic_qa/
Group / mailing list: epic-qa@list.nist.gov
- Recognizing Ultra Fine-grained Entities (RUFES)**
 The goal of the KBP RUFES track is to extract and corefer mentions of entities.
Track coordinator: Heng Ji (hengji@illinois.edu) and Avirup Sil (avirup@illinois.edu)
Home page: <https://tac.nist.gov/2020/KBP/RUFES/>
Group / mailing list: tac-kbp@list.nist.gov
- Streaming Multimedia Knowledge Base Population (SM-KBP)**
 The goal of the SM-KBP track is to develop and evaluate technologies for extracting and populating knowledge bases from streaming multimedia content.
Track coordinator: Hoa Dang (hoa.dang@nist.gov)
Home page: <https://tac.nist.gov/2020/KBP/SM-KBP/>
Group / mailing list: sm-kbp@list.nist.gov

<https://tac.nist.gov/2020/index.html>



De nombreux ateliers et compétitions (3)

DEFT (Défi Fouille de Textes)

- **2005** (*Dourdan, France, TALN 2005*) : identification du locuteur d'un discours politique parmi deux protagonistes différents (Jacques Chirac vs. François Mitterrand).
- **2006** (*Fribourg, Suisse, SDN 2006*) : segmentation thématique de textes politiques.
- **2007** (*Grenoble, France, AFIA 2007*) : détection de l'opinion exprimée dans un texte de retranscription de débats parlementaires (projets de Loi relatifs à l'énergie).
- **2008** (*Avignon, France, TALN 2008*) : classification automatique de documents en genres (*journalistique vs. encyclopédiques*) et thèmes différents (*art, économie, littérature, politique internationale, politique nationale, problèmes de sociétés, sciences, sports, télévision*).
- **2009** (*Paris, France*) : fouille d'opinion (objectif/subjectif) en corpus multilingues (journaux et débats européens).
- **2010** (*Montréal, Canada, TALN 2010*) :
 - Variation diachronique (1800-1944) en corpus de presse française (*Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires, La Croix, Le Figaro*), identification de la décennie de publication d'un extrait d'article ;
 - Variation diatopique en corpus de presse française (*L'Est Républicain, Le Monde*) et québécoise (*La Presse, Le Devoir*).
- **2011** (*Montpellier, France, TALN 2011*) :
 - Variations diachroniques (1800-1944) en corpus de presse française (*Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires, La Croix, Le Figaro, La Presse, Le Temps*), identification de l'année de publication d'un extrait d'article ;
 - Appariements résumé/article scientifique de revue dans le domaine des Sciences Humaines et Sociales (Humanités).
- **2012** (*Grenoble, France, TALN 2012*) : identification automatique des mots-clés indexant le contenu d'articles scientifiques ayant paru en revues de Sciences Humaines et Sociales, avec l'aide de la terminologie des mots-clés (piste 1), sans terminologie (piste 2).
- **2013** (*Les Sables-d'Olonne, France, TALN 2013*) : identification du niveau de difficulté de réalisation d'une recette, identification du type de plat préparé, appariement d'une recette avec son titre, identification des ingrédients d'une recette.
- **2014** (*Marseille, France, TALN 2014*) : catégoriser le genre littéraire de courtes nouvelles, évaluer la qualité littéraire de ces nouvelles, déterminer si une œuvre fait consensus, déterminer la session scientifique dans laquelle un article de conférence a été présenté.
- **2015** (*Caen, France, TALN 2015*) : fouille d'opinion, de sentiment et d'émotion dans des messages postés sur Twitter.
- **2016** (*Paris, France, TALN 2016*) : indexation de documents scientifiques en français.
- **2017** (*Orléans, France, TALN 2017*) : fouille d'opinion dans des messages postés sur Twitter.
- **2018** (*Rennes, France, CORIA-TALN 2018*) : recherche d'information et analyse de sentiments dans des tweets sur les transports en Ile-de-France.
- **2019** (*Toulouse, France, PFIA-TALN-RECITAL 2019*) : recherche
- **2020** (*Nancy, France, conférence virtuelle JEP-TALN-RECITAL*) : classification de cas cliniques

DEFT 2021

Défi Fouille de Textes@TALN 2021

Classification de cas cliniques et correction automatique de copies d'étudiants

<https://deft.lisn.upsaclay.fr/2021/>

