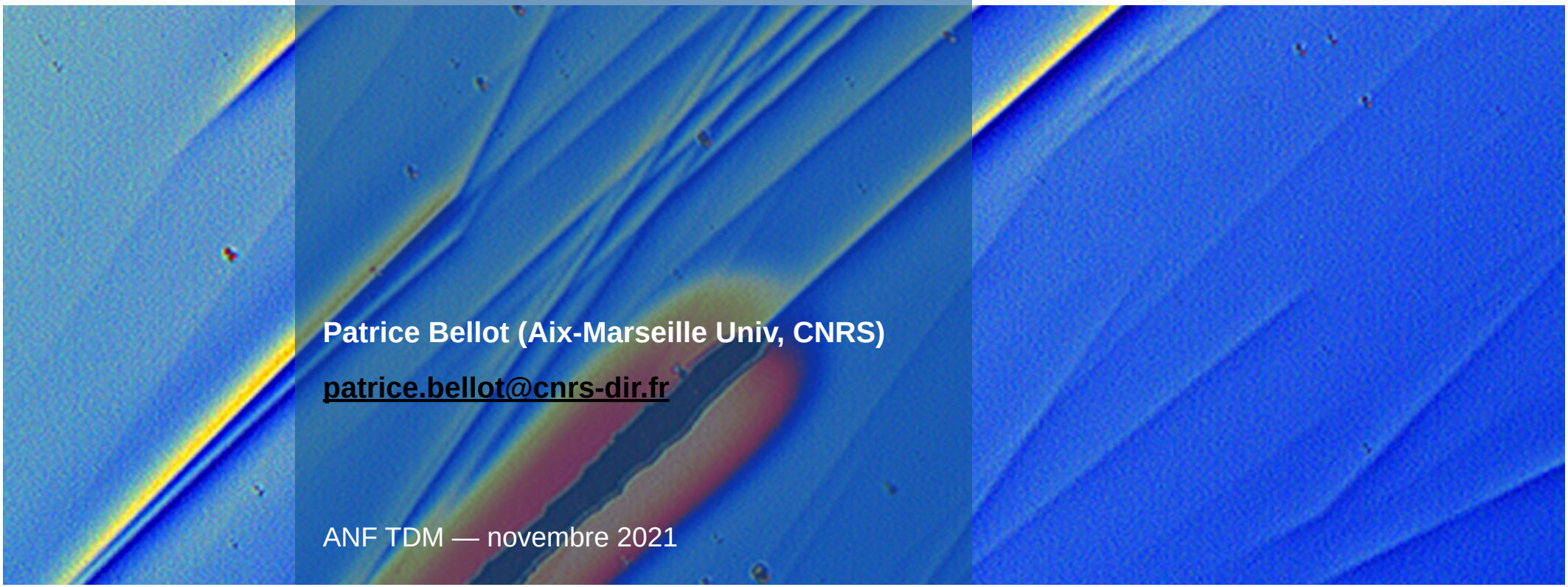
Introduction à la FOUILLE DE TEXTES et positionnement de l'offre logicielle



Patrice Bellot (Aix-Marseille Univ, CNRS)

patrice.bellot@cnrs-dir.fr

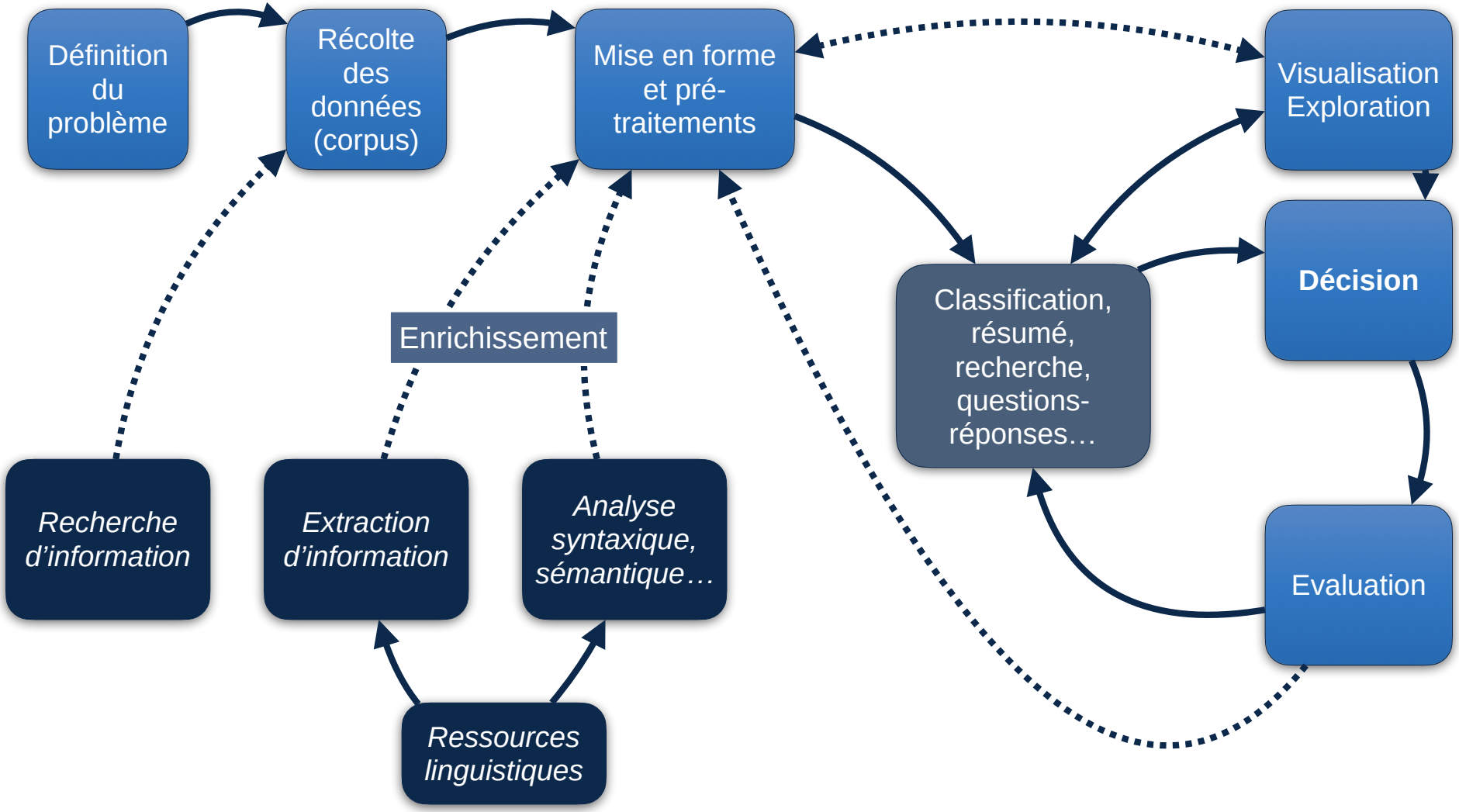
ANF TDM — novembre 2021

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 : lexique et terminologie, syntaxe…)
- La langue est ambiguë : polysémie, contexte, interprétation…

Des difficultés génériques.

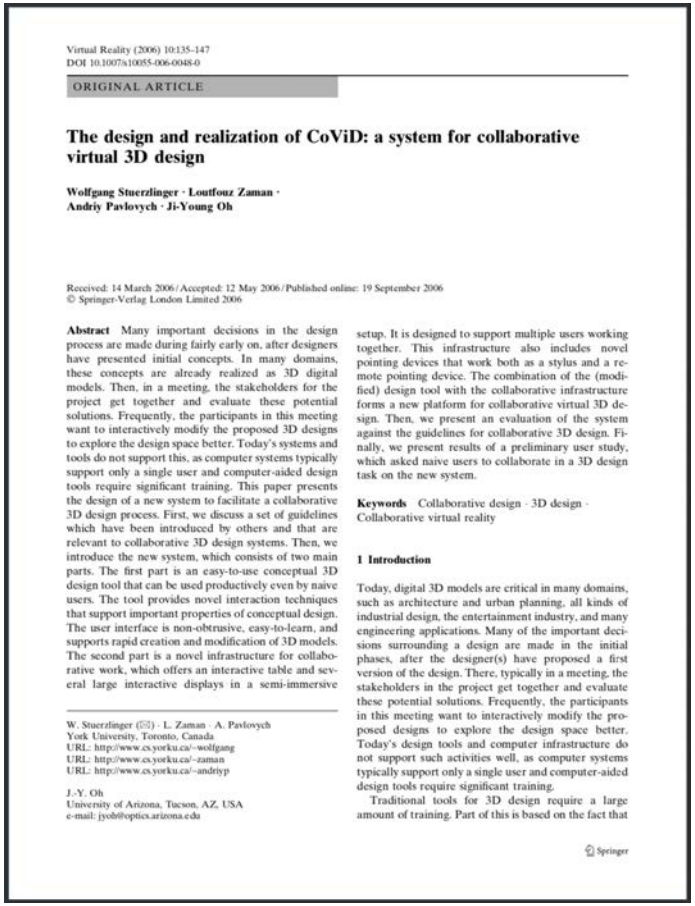
Des standards nécessaires.

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 · Loutfouz Zaman · Andriy Pavlovych · Ji-Young Oh

Received: 14 March 2006 / Accepted: 12 May 2006 / Published online: 19 September 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/~andriyp

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

Springer

=

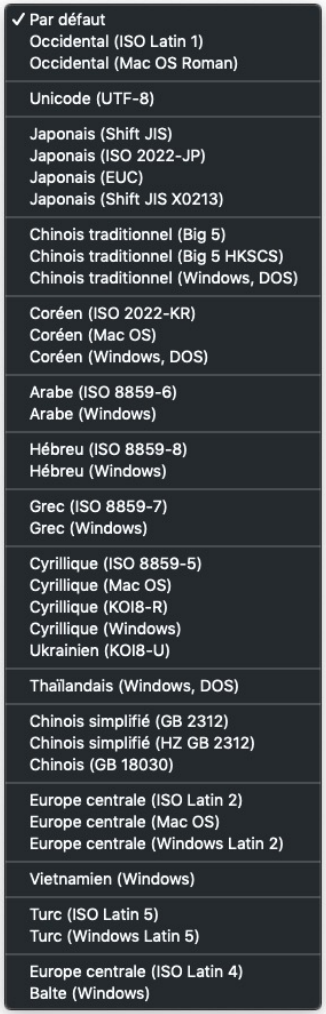
```
(base) [ Patrice Mac-Pro-de-Patrice ~ ] more /Users/Patrice/Downloads/ark__67375_V
"/Users/Patrice/Downloads/ark__67375_VQC-MZJ5GK10-Q/PDF/10055_2006_Article_48.pdf"
%PDF-1.3^M%<80><84><88><8C><90><94><98><9C><A0><A4><A8><AC><B0><B4><B8><BC><C0>
<C4><C8><CC><D0><D4><D8><DC><E0><E4><E8><EC><F0><F4><F8><FC>^M1 0 obj^M<< /CropBox
nots^M[ 2 0 R 3 0 R 4 0 R 5 0 R 6 0 R 7 0 R 8 0 R^M] /Rotate 0 /MediaBox^M[ 0 0 59
 0 R /Contents 10 0 R /Type /Page^M>>^Mendobj^M2 0 obj^M<< /Rect^M[ 367.483032 417
ype /Link /Border^M[ 0 0 0^M] /Type /Annot^M>>^Mendobj^M3 0 obj^M<< /Rect^M[ 435.6
R18) /Subtype /Link /Border
^M[ 0 0 0^M] /Type /Annot^M>>^Mendobj^M4 0 obj^M<< /Rect^M[ 306.142029 255.628006
k /Border^M[ 0 0 0^M] /Type /Annot^M>>^Mendobj^M5 0 obj^M<< /Rect^M[ 430.356018 15
type /Link /Border^M[ 0 0 0^M] /Type /Annot^M>>^Mendobj^M6 0 obj^M<< /Rect^M[ 454.
CR9) /Subtype /Link /Border^M[ 0 0 0^M] /Type /Annot^M>>^Mendobj^M7 0 obj^M<< /Rec
 /Dest (lCR10) /Subtype /Link /Border^M[ 0 0 0^M] /Type /Annot^M>>^Mendobj^M8 0 ob
.997009^M] /Dest (lCR6) /Subtype /Link /Border^M[ 0 0 0^M] /Type /Annot^M>>^Mendob
 185 0 R^M>> /ProcSet^M[ /PDF /Text^M] /ExtGState^M<< /GS1 163 0 R^M>> /ColorSpace
0 obj^M<< /Filter /FlateDecode /Length 6236^M>>^Mstream
H<89>tWM<8F><E3><B8>^Q<BD><F7><8F>^X87u<D0><ED><91>H<EA>++<B7>I^V<9B>_<97>^D X^C9
<EC>Ø<E8>ŠwdÑ<D1>^Q%[<9D>^?<B1>[<C9><DF>M<BD>*<CA>vQ^T^L<D0>CS<FC>(V<BD><F7><AA>s
10055_2006_Article_48.pdf
%PDF-1.3^M%<80><84><88><8C><90><94><98><9C><A0><A4><A8><AC><B0><B4><B8><BC><C0>
<C4><C8><CC><D0><D4><D8><DC><E0><E4><E8><EC><F0><F4><F8><FC>^M1 0 obj^M<< /CropBox
nots^M[ 2 0 R 3 0
 0 R /Contents 10
ype /Link /Border
R18) /Subtype /Li
 /Dest (lCR5) /Sub
997009^M] /Dest (l
5029 165.997009^M
12 490.109985 165.
9 156.019012 530.
 /F8 79 0 R /F2 1
 183 0 R^M>>
^M>>^Mendobj^M10 0
H<89>tWM<8F><E3><B
<EC>Ø<E8>ŠwdÑ<D1>^
<EA>_O<9F><FF>n<C
:...skipping...
%PDF-1.3^M%<80><84
8><FC>^M1 0 obj^M<<
Rotate 0 /MediaBox
```

The design and realization of CoViD: a system for collaborative virtual 3D design ISTEX Springer-Verlag London Springer-Verlag L
The design and realization of CoViD: a system for collaborative virtual 3D design Wolfgang Stuerzlinger York University Toronto C
CANADA Andriy Pavlovych York University Toronto CANADA Ji-Young Oh jyoh@optics.arizona.edu University of Arizona Tucso
D569A70620C3227A0BF16EF149FE0E7980229167 ark:/67375/VQC-MZJ5GK10-Q s10055-006-0048-0 10.1007/s10055-006-0048
4338 1434-9957 Springer-Verlag London 2006 10 2 135 147 pub2TEI-ISTEX A set of style sheets for converting XML documents er
common TEI format. We use TEI Abstract Many important decisions in the design process are made during fairly early on, after desig
domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get together an
participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today’s systems
typically support only a single user and computer-aided design tools require significant training. This paper presents the design of a n
process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D design sy
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 user
support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rapid creation and
infrastructure for collaborative work, which offers an interactive table and several large interactive displays in a semi-immersive setu
together. This infrastructure also includes novel pointing devices that work both as a stylus and a remote pointing device. The combin
collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present an evaluation of the system ag
Finally, we present results of a preliminary user study, which asked naive users to collaborate in a 3D design task on the new system.
Vision Artificial Intelligence (incl. Robotics) Computing Methodologies Computer Graphics Collaborative design 3D design Collabor
Accepted ePublished Virtual Reality (2006) 10:135–147 DOI 10.1007/s10055-006-0048-0 O R I G I N A L A R T I C L E The design
virtual 3D design Wolfgang Stuerzlinger Æ Loutfouz Zaman Æ Andriy Pavlovych Æ Ji-Young Oh Received: 14 March 2006 / Accep
2006 Springer-Verlag London Limited 2006 Abstract Many important decisions in the design process are made during fairly early on,
many domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get togetl
the participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today’s syste
systems typically support only a single user and computer-aided design tools require significant training. This paper presents the desig
design process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D de
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 nai
techniques that support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rap
second part is a novel infrastructure for collaborative work, which offers an interactive table and several large interactive displays in a
A. Pavlovych York University, Toronto, Canada URL: http://www.cs.yorku.ca/~wolfgang URL: http://www.cs.yorku.ca/~zaman URL
University of Arizona, Tucson, AZ, USA e-mail: jyoh@optics.arizona.edu setup. It is designed to support multiple users working tog
devices that work both as a stylus and a remote pointing device. The combination of the (modified) design tool with the collaborative
virtual 3D design. Then, we present an evaluation of the system against the guidelines for collaborative 3D design. Finally, we presen
naive users to collaborate in a 3D design task on the new system. Keywords Collaborative design 3D design Collaborative virtual real
critical in many domains, such as architecture and urban planning, all kinds of industrial design, the entertainment industry, and many
decisions surrounding a design are made in the initial phases, after the designer(s) have proposed a first version of the design. There, t
get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the propos

PDF to text

Evitons les PDF…

L’encodage des caractères

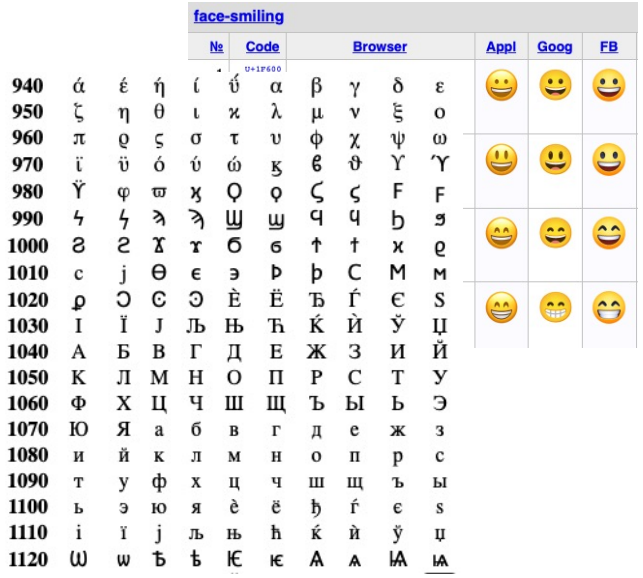


Dec	Hex	Char	Dec	Hex	Char
64	40	@	96	60	`
65	41	A	97	61	a
66	42	B	98	62	b
67	43	C	99	63	c
68	44	D	100	64	d
69	45	E	101	65	e
70	46	F	102	66	f
71	47	G	103	67	g
72	48	H	104	68	h
73	49	I	105	69	i
74	4A	J	106	6A	j
75	4B	K	107	6B	k
76	4C	L	108	6C	l
77	4D	M	109	6D	m
78	4E	N	110	6E	n
79	4F	O	111	6F	o
80	50	P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

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

7	00A8	00A9	00AA	00AB	00AC	00A
7	00B8	00B9	00BA	00BB	00BC	00B
7	00C8	00C9	00CA	00CB	00CC	00C
7	00D8	00D9	00DA	00DB	00DC	00D
7	00E8	00E9	00EA	00EB	00EC	00E
7	00F8	00F9	00FA	00FB	00FC	00F

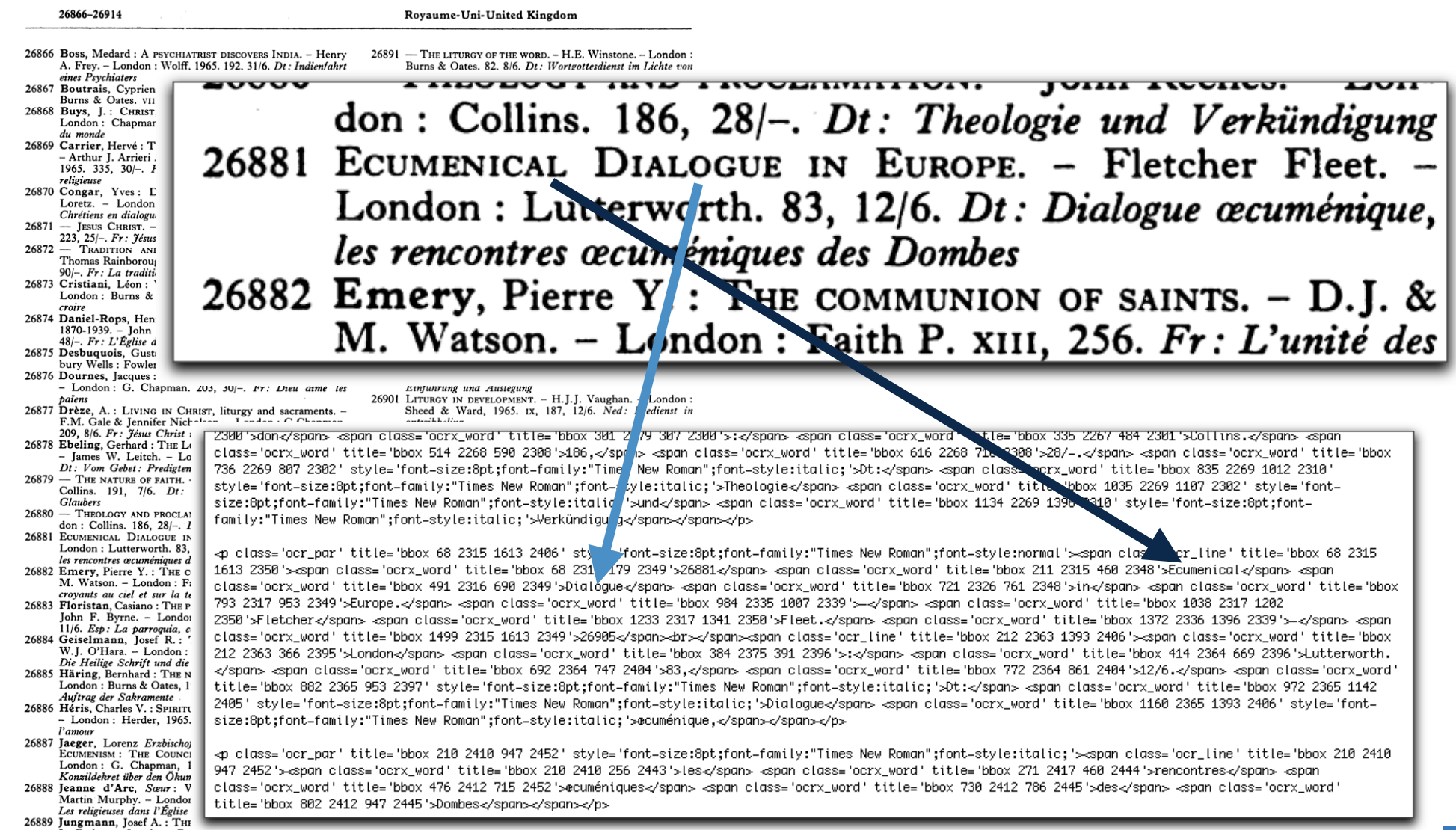
ISO-Latin 1 sur 8 bits
(ISO/CEI 8859)



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

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

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. Arrieri . 1965. 335, 30/–. *religieuse*

26870 **Congar**, Yves : L Loretz. – London *Chrétiens en dialogu*

26871 — JESUS CHRIST. – 223, 25/–. *Fr : Jésus*

26872 — TRADITION ANI Thomas Rainborou 90/–. *Fr : La traditi*

26873 **Cristiani**, Léon : London : Burns & *croire*

26874 **Daniel-Rops**, Hen 1870-1939. – John 48/–. *Fr : L'Église a*

26875 **Desbuquois**, Gust bury Wells : Fowle

26876 **Dournes**, Jacques : – London : G. Chapman. 203, 30/–. *Fr : Dieu aime les païens*

26877 **Drèze**, A. : LIVING IN CHRIST, liturgy and sacraments. – F.M. Gale & Jennifer Nich 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 PROCLA don : Collins. 186, 28/–.

26881 ECUMENICAL DIALOGUE IN London : Lutterworth. 83, *les rencontres œcuméniques d*

26882 **Emery**, Pierre Y. : THE C M. Watson. – London : F *croyants au ciel et sur la t*

26883 **Floristan**, Casiano : THE P John F. Byrne. – Londo 11/6. *Esp : La parroquia, c*

26884 **Geiselmann**, Josef R. : W.J. O'Hara. – London : *Die Heilige Schrift und die*

26885 **Häring**, Bernhard : THE N London : Burns & Oates, 1 *Auftrag der Sakramente*

26886 **Héris**, Charles V. : SPIRITU – London : Herder, 1965. *l'amour*

26887 **Jaeger**, Lorenz *Erzbischo* ECUMENISM : THE COUNCI London : G. Chapman, 1 *Konzildekret über den Ökun*

26888 **Jeanne d'Arc**, *Sœur* : V Martin Murphy. – Londo *Les religieuses dans l'Église*

26889 **Jungmann**, Josef A. : THI 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*

don : Collins. 186, 28/–. *Dt : Theologie und Verkündigung*

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*

Einführung und Auslegung

26901 LITURGY IN DEVELOPMENT. – H.J.J. Vaughan. – London : Sheed & Ward, 1965. IX, 187, 12/6. *Ned : Liedienst in*


```
2300'>don</span> <span class='ocrx_word' title='bbox 301 2279 307 2300'>:</span> <span class='ocrx_word' title='bbox 335 2267 484 2301'>Collins.</span> <span
class='ocrx_word' title='bbox 514 2268 590 2308'>186,</span> <span class='ocrx_word' title='bbox 616 2268 71[illegible]308'>28/-.</span> <span class='ocrx_word' title='bbox
736 2269 807 2302' style='font-size:8pt;font-family:"Time[illegible] New Roman";font-style:italic;'>Dt:</span> <span class='ocrx_word' title='bbox 835 2269 1012 2310'
style='font-size:8pt;font-family:"Times New Roman";font-style:italic;'>Theologie</span> <span class='ocrx_word' title='bbox 1035 2269 1107 2302' style='font-
size:8pt;font-family:"Times New Roman";font-style:italic;'>und</span> <span class='ocrx_word' title='bbox 1134 2269 1396 2310' style='font-size:8pt;font-
family:"Times New Roman";font-style:italic;'>Verkündigung</span></span></p>

<p class='ocr_par' title='bbox 68 2315 1613 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 460 2348'>Ecumenical</span> <span
class='ocrx_word' title='bbox 491 2316 690 2349'>Dialogue</span> <span class='ocrx_word' title='bbox 721 2326 761 2348'>in</span> <span class='ocrx_word' title='bbox
793 2317 953 2349'>Europe.</span> <span class='ocrx_word' title='bbox 984 2335 1007 2339'>–</span> <span class='ocrx_word' title='bbox 1038 2317 1202
2350'>Fletcher</span> <span class='ocrx_word' title='bbox 1233 2317 1341 2350'>Fleet.</span> <span class='ocrx_word' title='bbox 1372 2336 1396 2339'>–</span> <span
class='ocrx_word' title='bbox 1499 2315 1613 2349'>26905</span><br></span><span class='ocr_line' title='bbox 212 2363 1393 2406'><span class='ocrx_word' title='bbox
212 2363 366 2395'>London</span> <span class='ocrx_word' title='bbox 384 2375 391 2396'>:</span> <span class='ocrx_word' title='bbox 414 2364 669 2396'>Lutterworth.
</span> <span class='ocrx_word' title='bbox 692 2364 747 2404'>83,</span> <span class='ocrx_word' title='bbox 772 2364 861 2404'>12/6.</span> <span class='ocrx_word'
title='bbox 882 2365 953 2397' style='font-size:8pt;font-family:"Times New Roman";font-style:italic;'>Dt:</span> <span class='ocrx_word' title='bbox 972 2365 1142
2405' style='font-size:8pt;font-family:"Times New Roman";font-style:italic;'>Dialogue</span> <span class='ocrx_word' title='bbox 1160 2365 1393 2406' style='font-
size:8pt;font-family:"Times New Roman";font-style:italic;'>œcuménique,</span></span></p>

<p class='ocr_par' title='bbox 210 2410 947 2452' style='font-size:8pt;font-family:"Times New Roman";font-style:italic;'><span class='ocr_line' title='bbox 210 2410
947 2452'><span class='ocrx_word' title='bbox 210 2410 256 2443'>les</span> <span class='ocrx_word' title='bbox 271 2417 460 2444'>rencontres</span> <span
class='ocrx_word' title='bbox 476 2412 715 2452'>œcuméniques</span> <span class='ocrx_word' title='bbox 730 2412 786 2445'>des</span> <span class='ocrx_word'
title='bbox 802 2412 947 2445'>Dombes</span></span></p>
```


26914 Roman Catholic Church. *2d Vatican Ecumenical Council.* DECLARATION ON RELIGIOUS LIBERTY. – Thomas Atthill. – London : Catholic Truth Society. 19, 1/–. *Lat : De libertate religiosa*

567

Des balises orientées affichage

BACHELARD, Gaston

1975 *La formation de l'esprit scientifique* (Paris, Librairie philosophique J. Vrin) [1re 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">A</hi><hi rend="small-caps bold">mossy</hi><hi rend="bold">, Ruth<lb/></hi>1997 <hi rend=
    <bibl><hi rend="bold">A</hi><hi rend="small-caps bold">ron</hi><hi rend="bold">, Raymond<lb/></hi>1967 <hi rend
    <bibl><hi rend="bold">A</hi><hi rend="small-caps bold">sséo</hi><hi rend="bold">, Henriette<lb/></hi>1974 Le tr
    <bibl><hi rend="bold">B</hi><hi rend="small-caps bold">achelard</hi><hi rend="bold">, Gaston<lb/></hi>1975 <hi
    <bibl><hi rend="bold">B</hi><hi rend="small-caps bold">arthes</hi><hi rend="bold">, Roland<lb/></hi>1957 <hi re
    <bibl><hi rend="bold">B</hi><hi rend="small-caps bold">enveniste</hi><hi rend="bold">, Émile<lb/></hi>1966 De l
    <bibl><hi rend="bold">C</hi><hi rend="small-caps bold">oupry</hi><hi rend="bold">, François<lb/></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<lb/></hi>1994 Rétr
    <bibl><hi rend="bold">D</hi><hi rend="small-caps bold">ubois</hi><hi rend="bold">, Jacques<lb/></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, Bozidzr Ed.).</hi>
    </bibl>
    <bibl><hi rend="bold">F</hi><hi rend="small-caps bold">erney</hi><hi rend="bold">, Alice<lb/></hi>1997 <hi rend
    <bibl><hi rend="bold">G</hi><hi rend="small-caps bold">eertz</hi><hi rend="bold">, Clifford<lb/></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.<lb/></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>
```

et des annotations sémantiques

XML

```
<?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">Spring
            <pubPlace>London</pubPlace>
            <availability>
               <licence>Springer-Verlag London Limited</licence>
               <p scheme="https://loaded-corpus.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 type="publication-type"
                  subtype="journal"
                  scheme="https://publication-type.data.istex.fr/ark:/67375/JMC-5WTPMB5N-F">journal</not
         </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>
```

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

VS

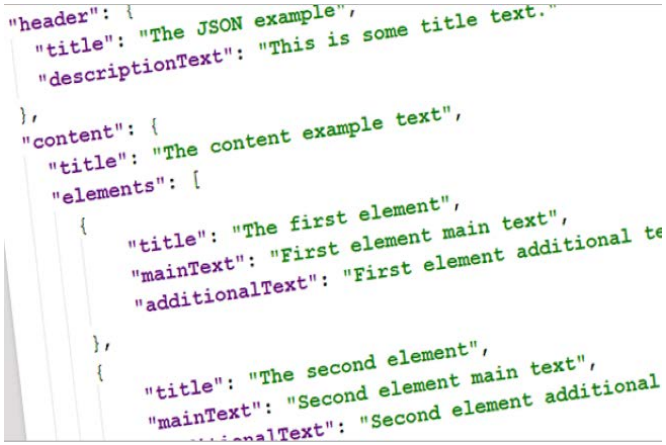
```
Abstract Many important decisions in the design
process are made during fairly early on, after designers
have presented initial concepts. In many domains,
these concepts are already realized as 3D digital
models. Then, in a meeting, the stakeholders for the
project get together and evaluate these 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;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 re
which asked naive users to collaborate in a 3D design
task on the new system.
Keywords Collaborative design  3D design
Collaborative virtual reality
```

De multiples formats autres que XML

```
"jour";"nomReg";"numReg";"incid_rea"
2020-03-19;"Auvergne-RhÛne-Alpes";84;44
2020-03-19;"Bourgogne-Franche-ComtÈ";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 RÈunion";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-CÛte d'Azur";93;25
2020-03-20;"Auvergne-RhÛne-Alpes";84;16
2020-03-20;"Bourgogne-Franche-ComtÈ";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



JSON et ses variantes

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

```yaml
---
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

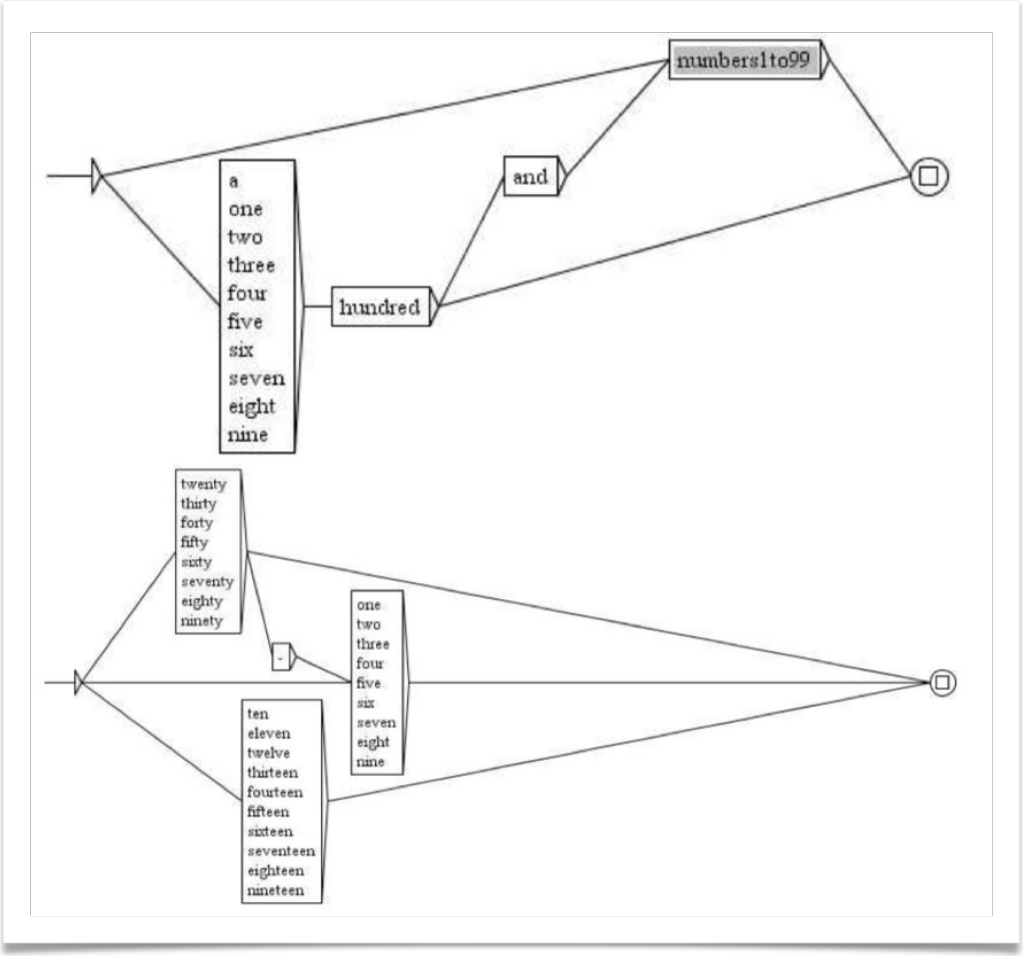
Yet Another Markup Language / YAML Ain't Markup Language

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

Des méthodes informatiques et des ressources linguistiques

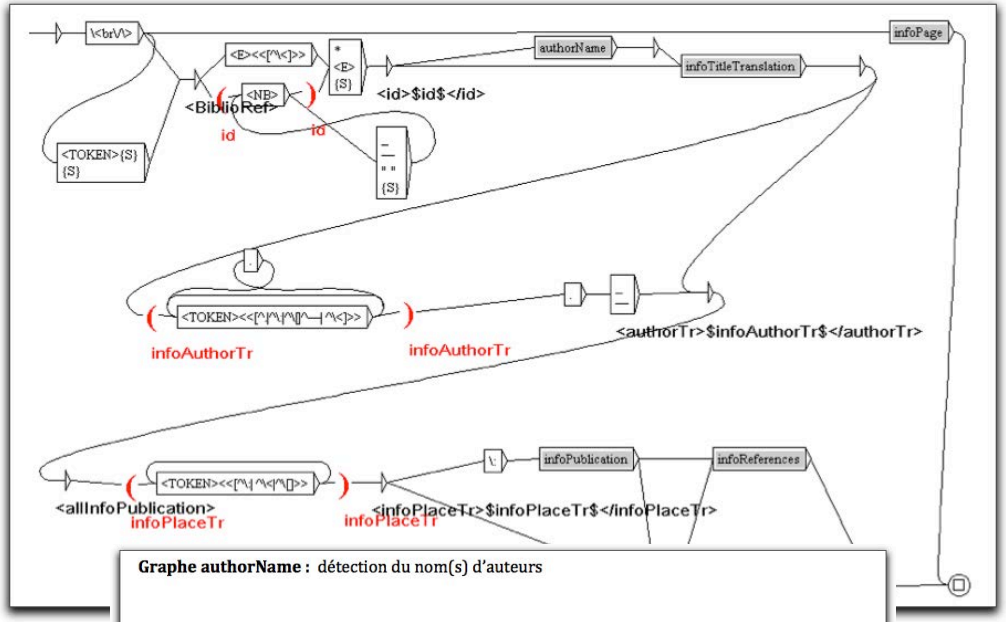
04.06.19

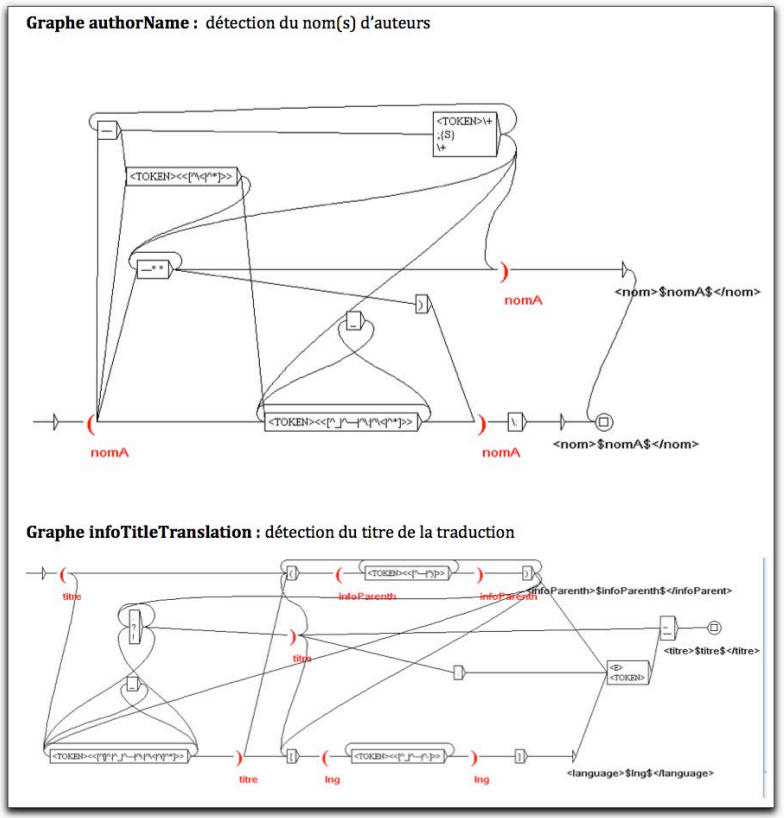
Des symboles et des automates



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

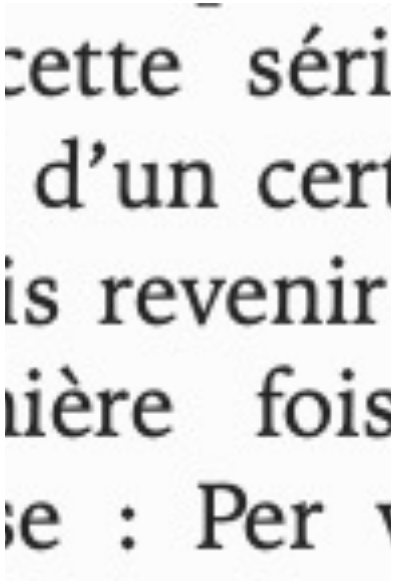
https://hal.archives-ouvertes.fr/hal-00145253/document



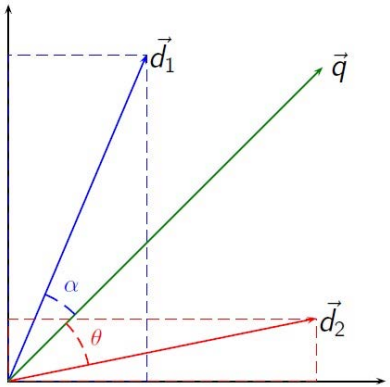


Etude pour l'UNESCO (Univ. Avignon, Open Edition 2011)

Symboles, vecteurs et modèles de langue



Un mot = une forme



Des vecteurs de mots

$$\text{Similarité}(d_1, d_2) \approx \overrightarrow{d_1}.\overrightarrow{d_2}$$

$$\text{Similarité}(d_1, d_2) \approx \cos(\overrightarrow{d_1}, \overrightarrow{d_2})$$

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

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

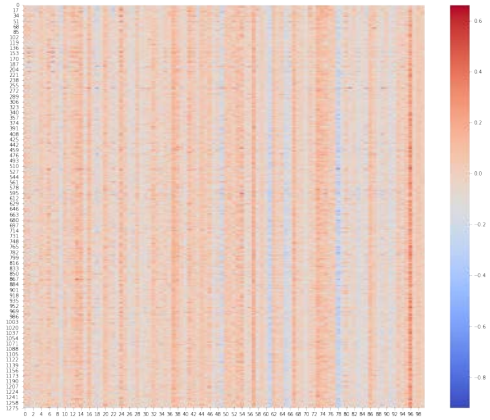
positive	Z_score	negative	Z_score	Neutral	Z_score
Love	14.31	Not	13.99	Httpbit	6.44
Good	14.01	Fuck	12.97	Httpfb	4.56
Happy	12.30	Don't	10.97	Httpbnd	3.78
Great	11.10	Shit	8.99	Intern	3.58
Excite	10.35	Bad	8.40	Nov	3.45
Best	9.24	Hate	8.29	Httpdlvr	3.40
Thank	9.21	Sad	8.28	Open	3.30
Hope	8.24	Sorry	8.11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

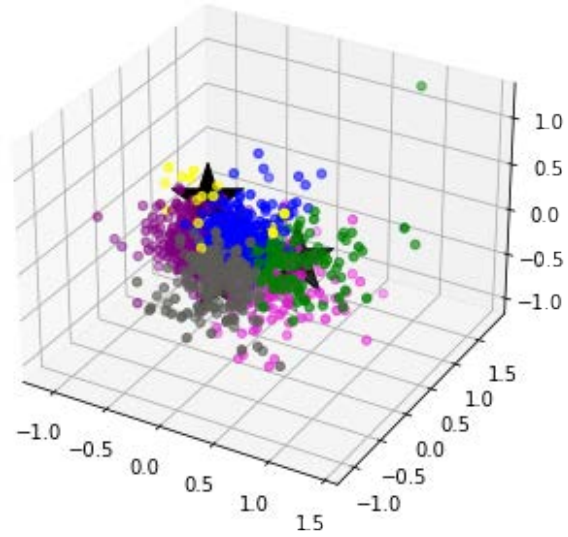
$$P(Classe|(w_1, w_2, ..., w_{T-1}, w_T))$$

et règle de Bayes

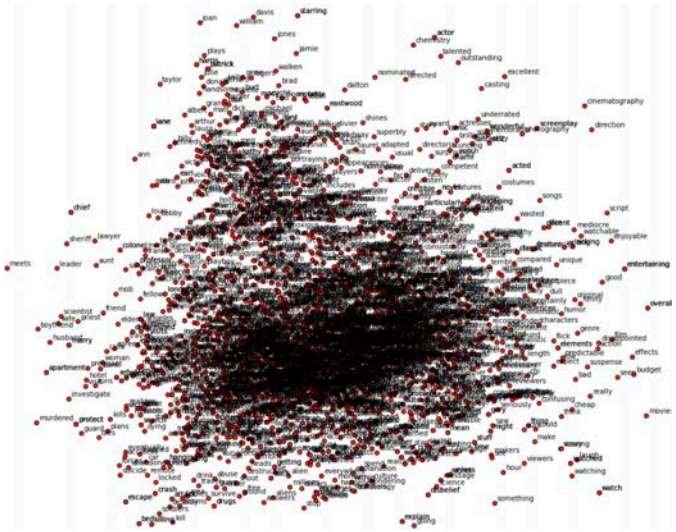
et des représentations

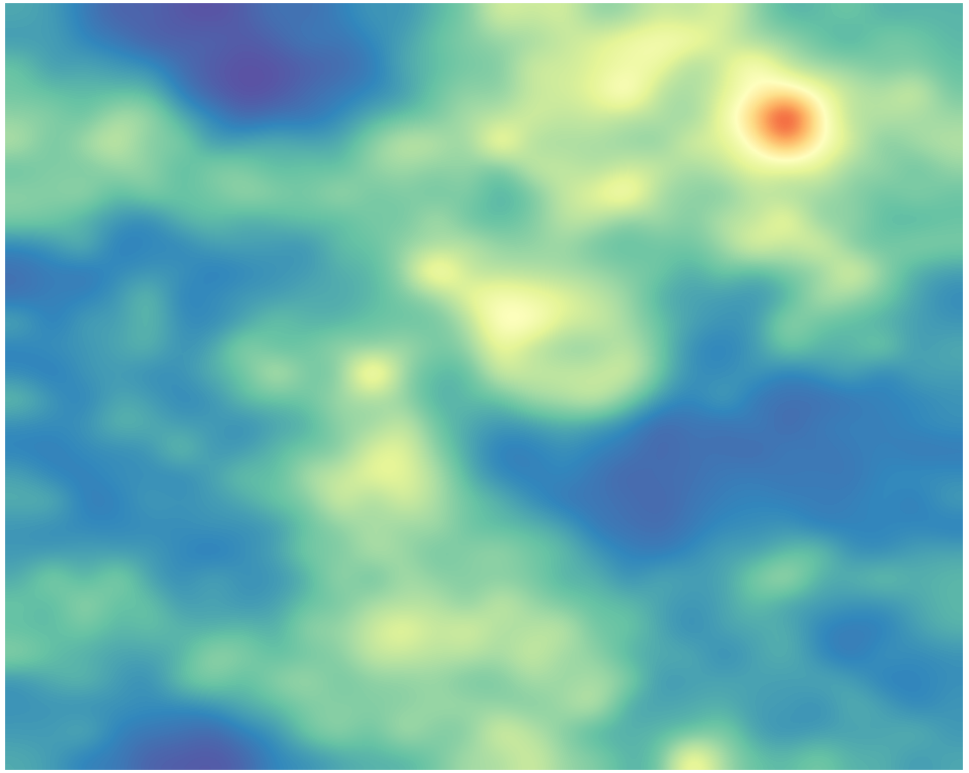


Vectorisation dans des espaces continus (Doc2Vec) par plongements lexicaux



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.

Google Books

Ngram Viewer Exports

The Google Books Ngram Viewer is optimized for quick inquiries into the usage of small sets of phrases. If you're interested i 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
```

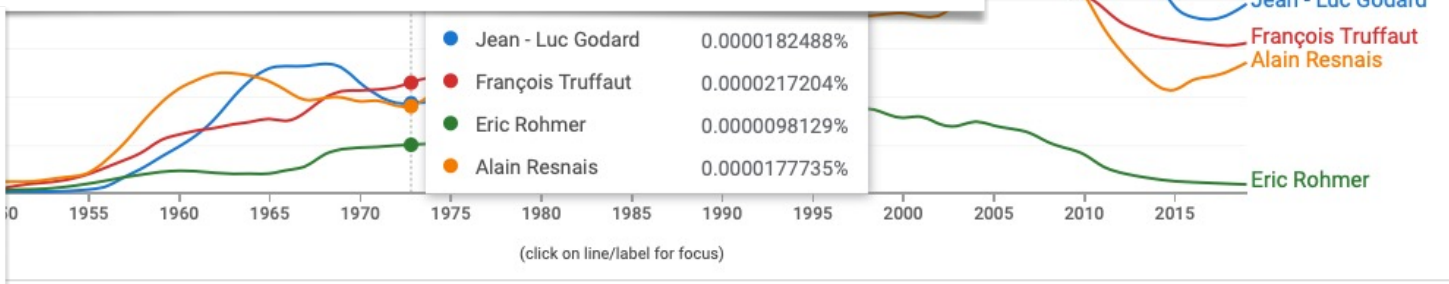
French

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

https://huggingface.co/transformers/model_summary.html

CamemBERT

A Tasty French Language Model

Facebook AI Research

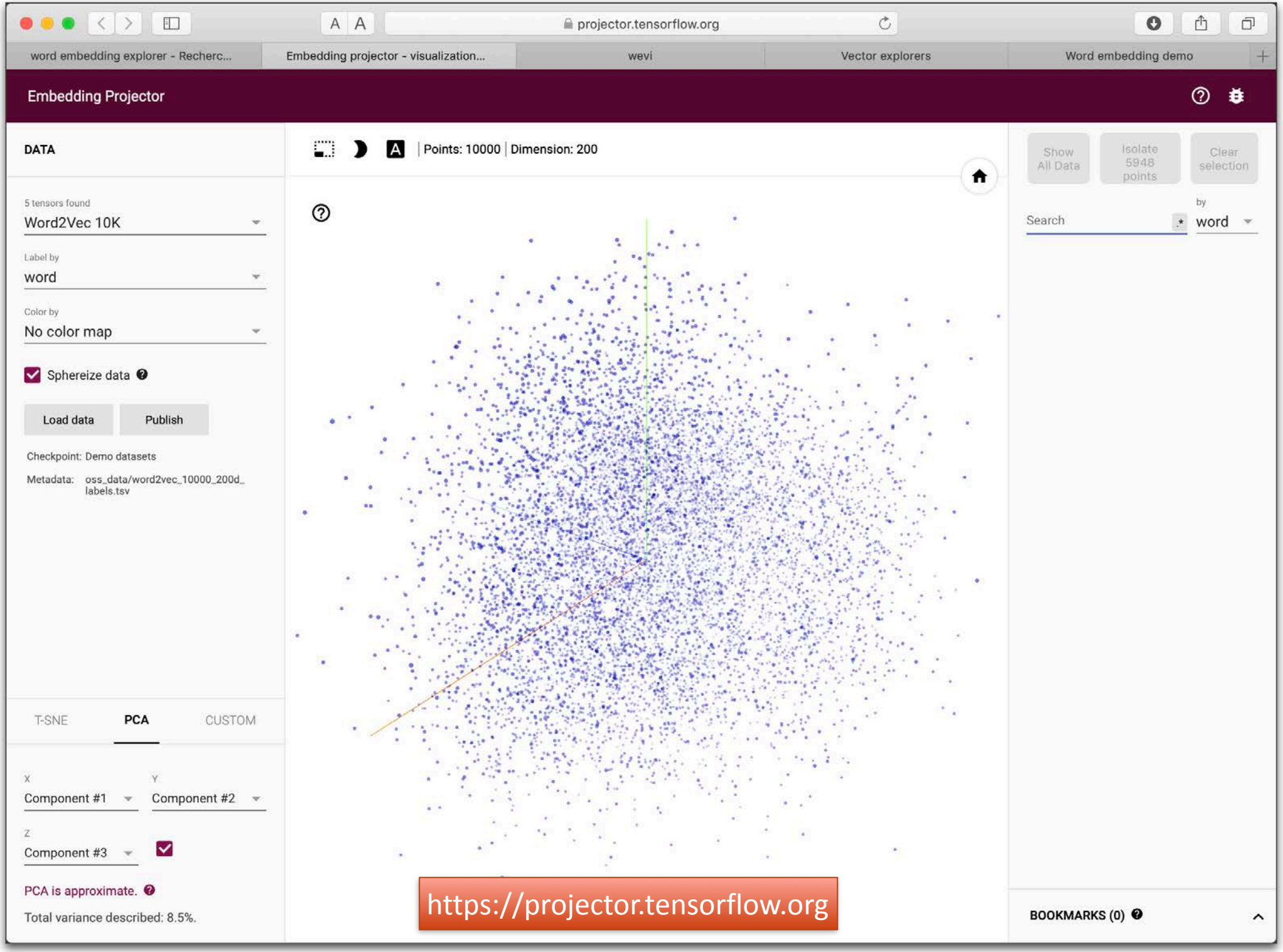
Inria

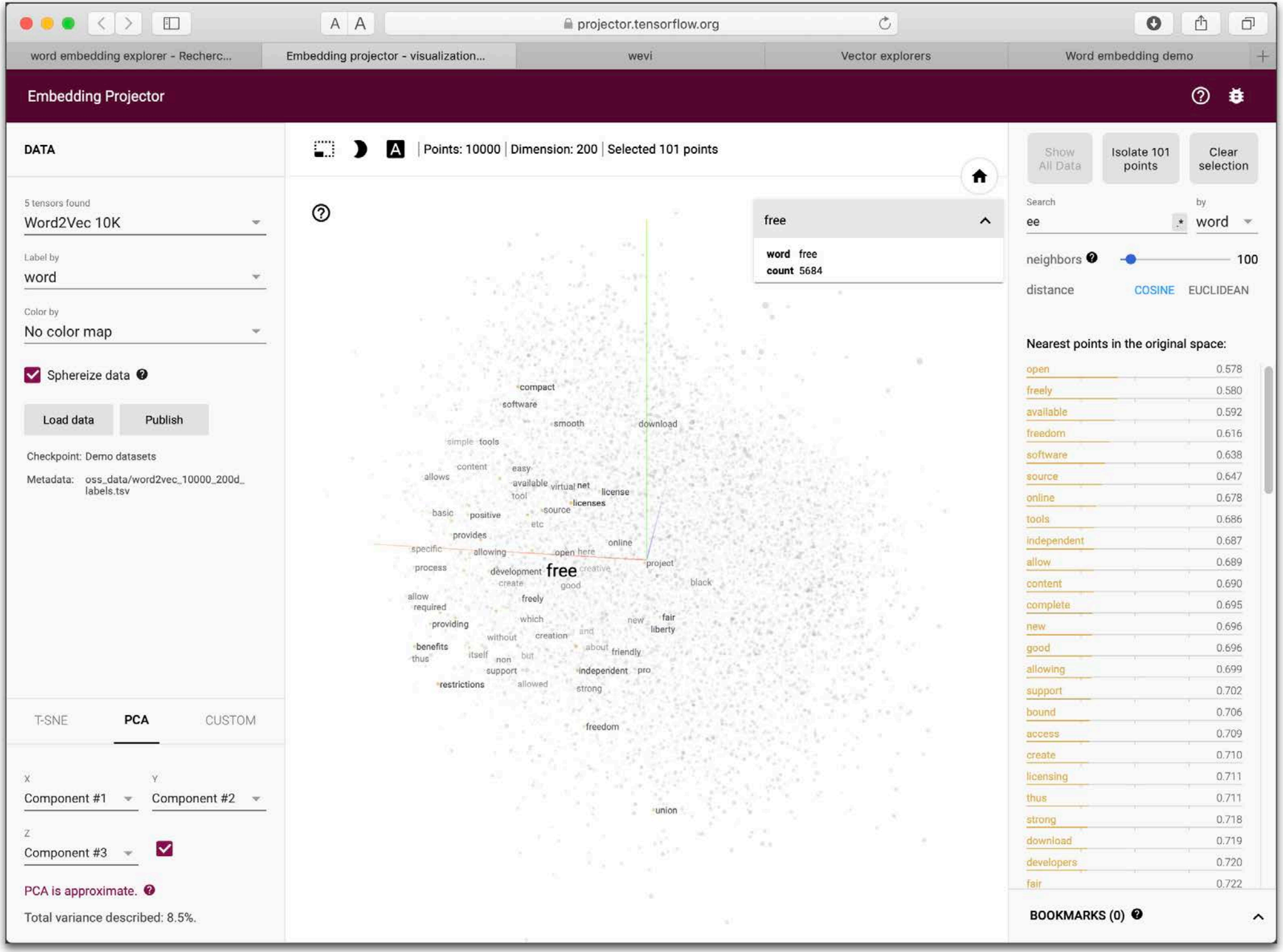
ALMAnaCH

https://camembert-model.fr

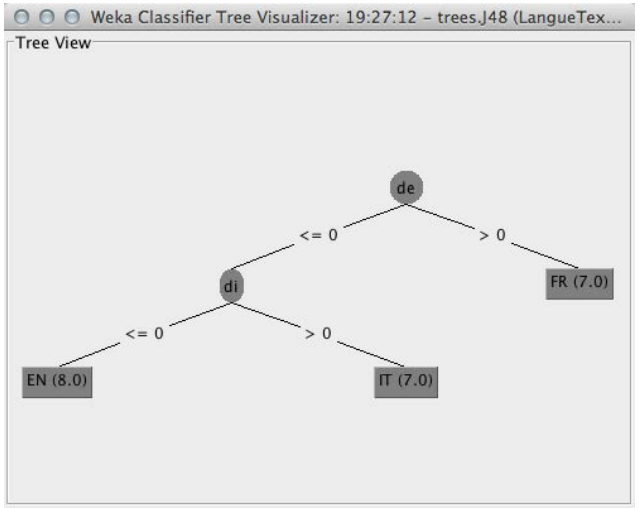
FlauBERT

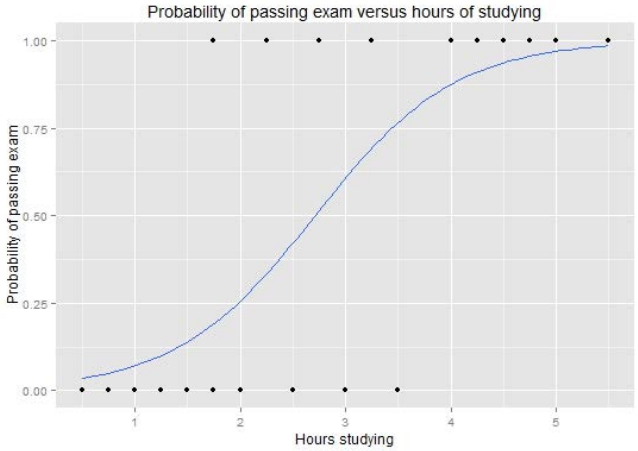
https://github.com/getalp/Flaubert


projector.tensorflow.org
word embedding explorer - Recherc...
Embedding projector - visualization...
wevi
Vector explorers
Word embedding demo
Embedding Projector
DATA
Points: 10000 | Dimension: 200
Show All Data
Isolate 5948 points
Clear selection
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
Search
by
word
T-SNE
PCA
CUSTOM
X
Component #1
Y
Component #2
Z
Component #3
PCA is approximate.
Total variance described: 8.5%.
https://projector.tensorflow.org
BOOKMARKS (0)

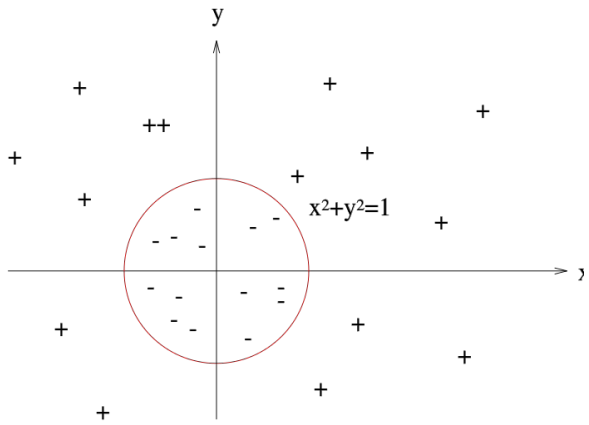

projector.tensorflow.org
word embedding explorer - Recherc...
Embedding projector - visualization...
wevi
Vector explorers
Word embedding demo
Embedding Projector
DATA
5 tensors found
Word2Vec 10K
Label by
word
Color by
No color map
Sphereize data
Load data
Publish
Checkpoint: Demo datasets
Metadata: 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 | Selected 101 points
free
word free
count 5684
compact
software
smooth
download
simple
tools
content
easy
allows
available
virtual net
tool
license
licenses
basic
positive
source
etc
provides
online
specific
allowing
open
here
process
development
free
creative
project
black
create
good
allow
required
freely
providing
which
new
fair
liberty
without
creation
and
benefits
thus
itself
non
but
about
friendly
support
independent
pro
restrictions
allowed
strong
freedom
union
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)

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

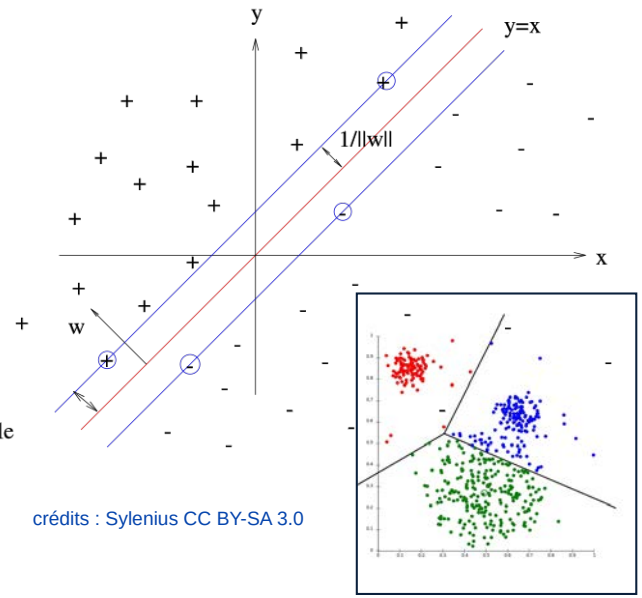




crédits : Michaelg2015 CC BY-SA 4.0

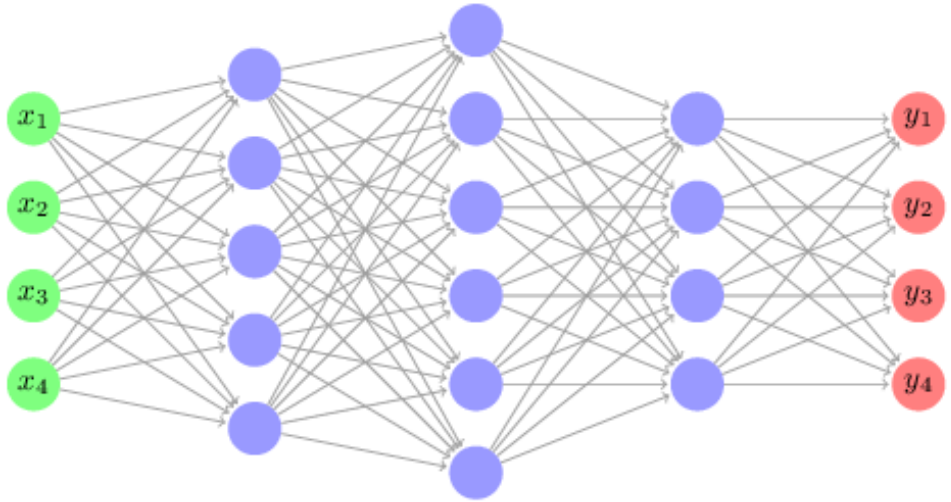


crédits : Sylenius CC BY-SA 3.0

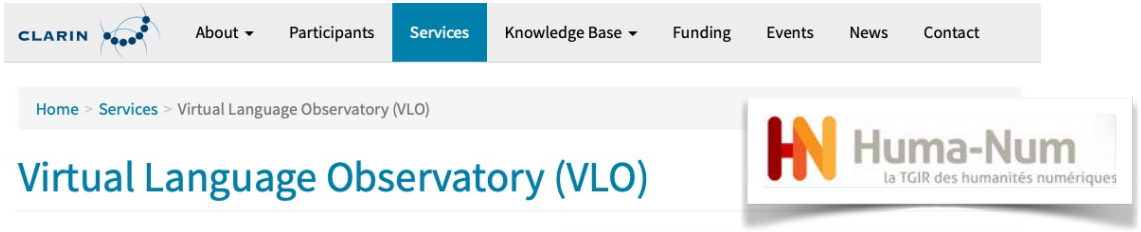


Marge maximale

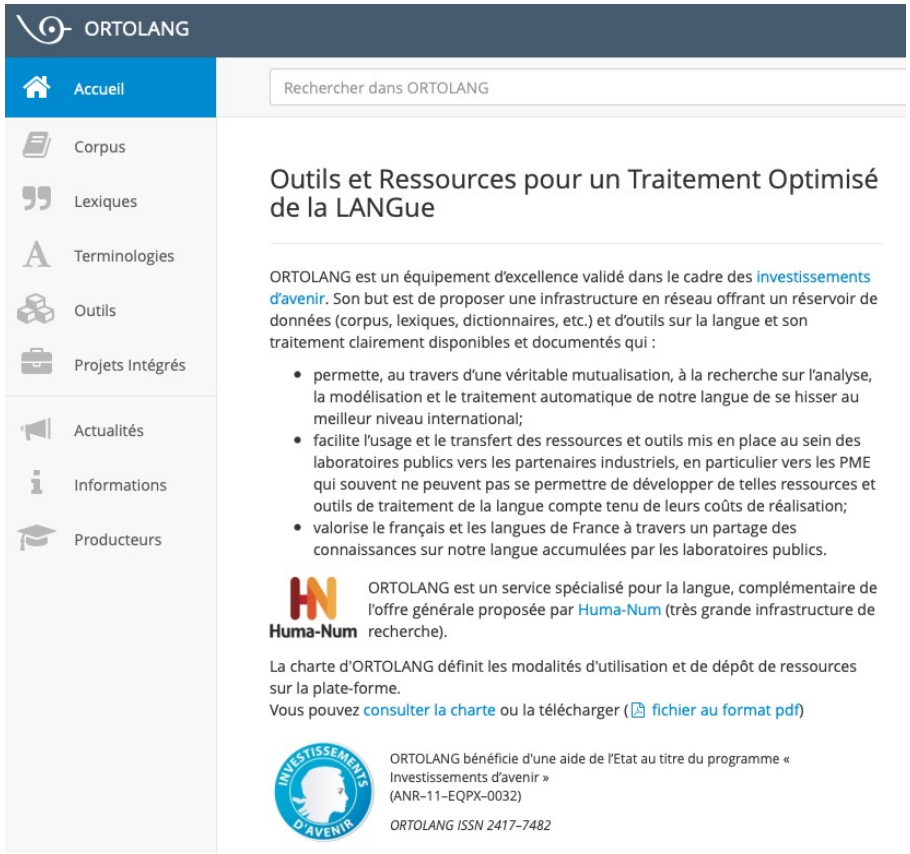
crédits : Sylenius CC BY-SA 3.0



données observées → Sorties avec scores



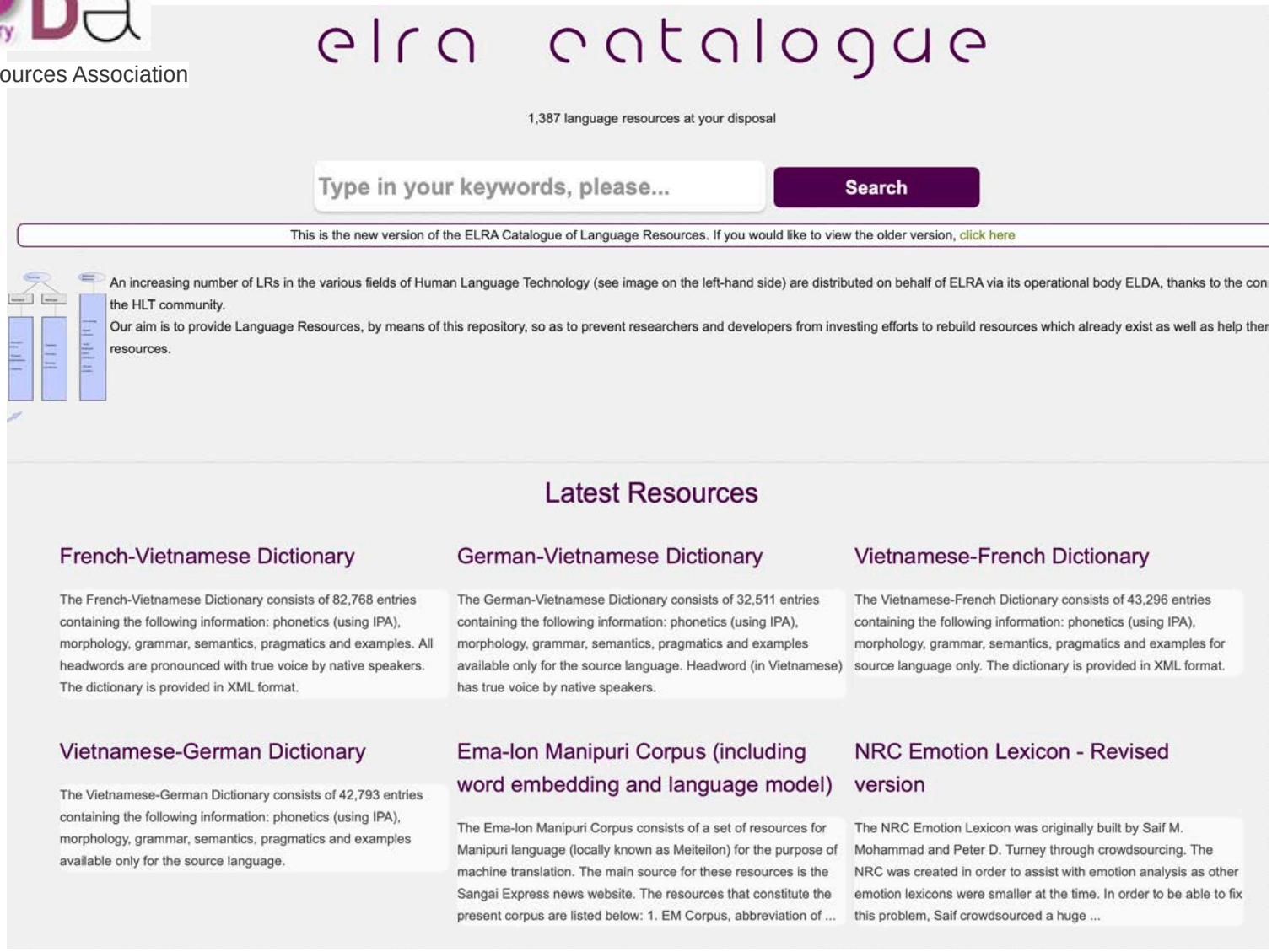
https://www.clarin.eu/content/virtual-language-observatory-vlo



https://www.ortolang.fr

cnrs

European Language Resources Association



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

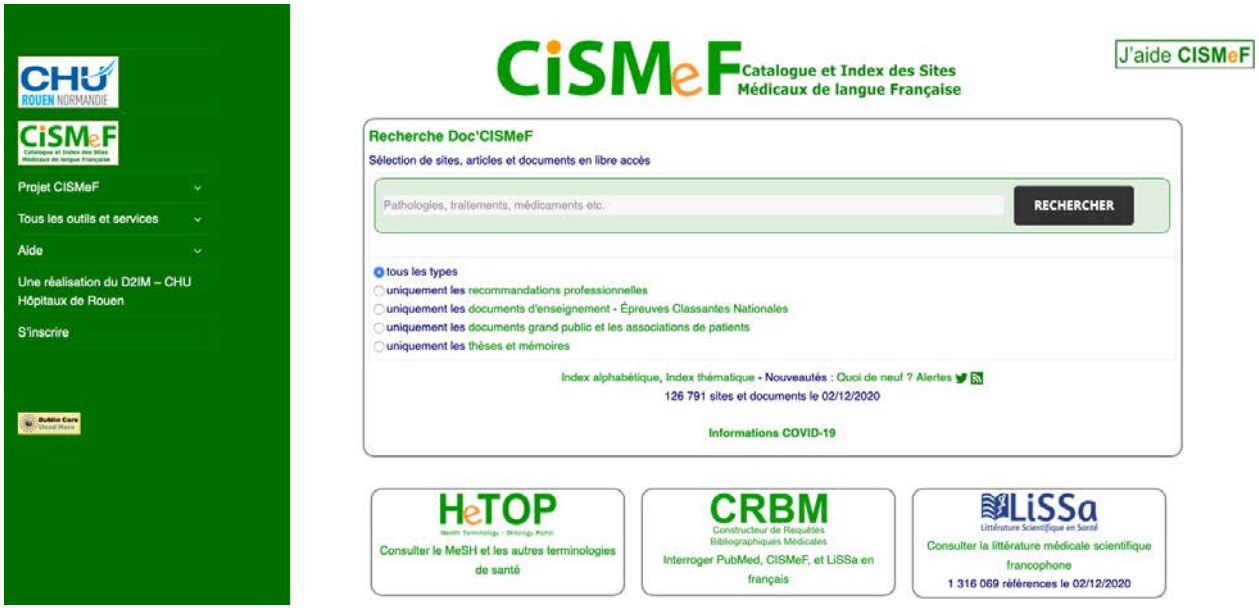
voir aussi : https://www.ldc.upenn.edu

Linguistic Data Consortium

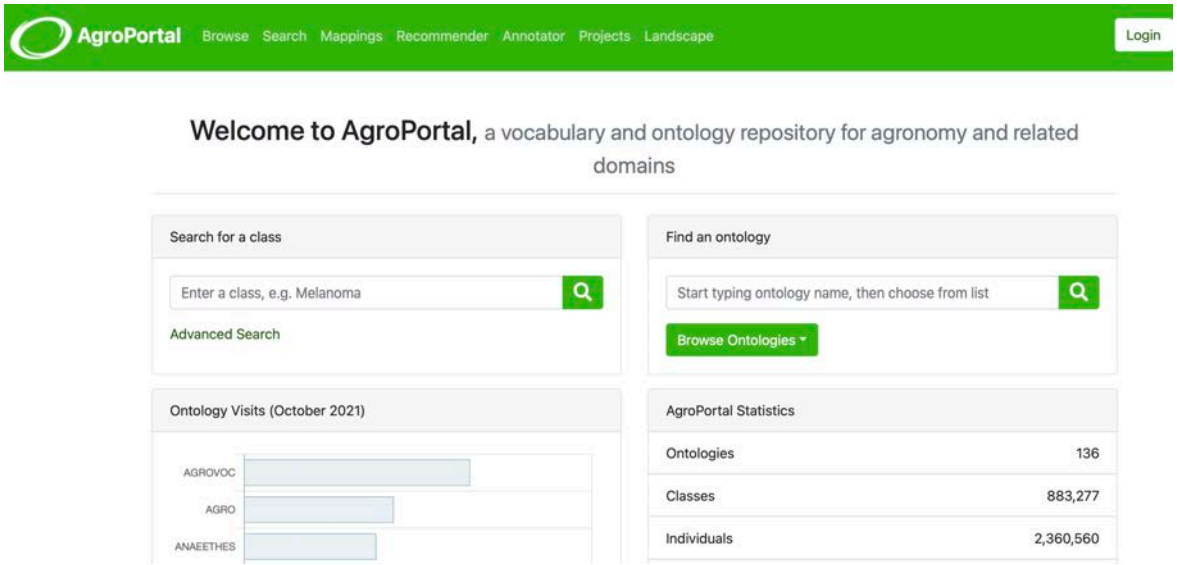
Ressources « collaboratives »

http://www.jeuxdemots.org

Et des ressources spécialisées



http://www.chu-rouen.fr/cismef/



http://agroportal.lirmm.fr

Microsoft | Research Our research Programs & events Blogs & podcasts About Sign up: Research Newsletter All Microsoft

Overview Publications

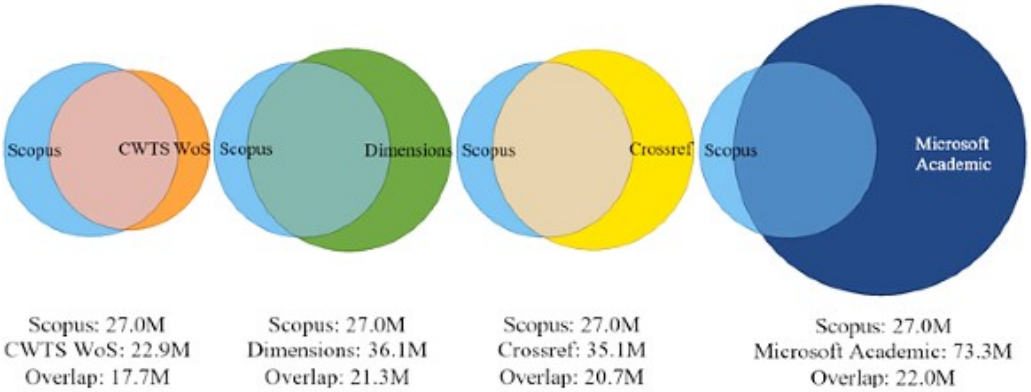
Open Academic Graph (OAG) is a large knowledge graph unifying two billion-scale academic graphs: Microsoft Academic Graph (MAG) and AMiner. In mid 2017, we published OAG v1, which contains 166,192,182 papers from MAG and 154,771,162 papers from AMiner (see below) and generated 64,639,608 linking (matching)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

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.



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

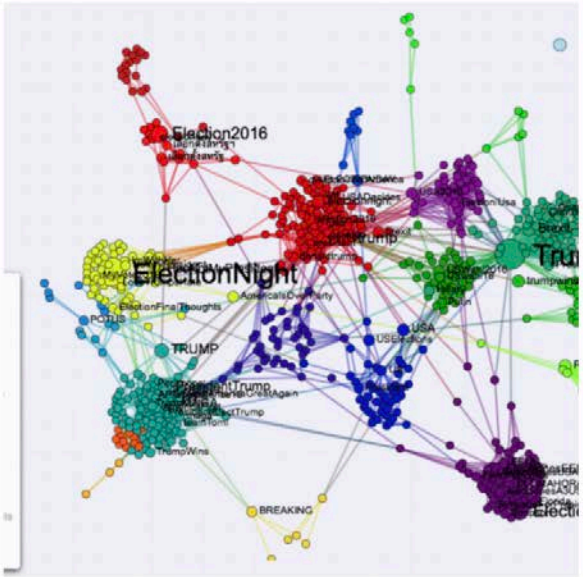
https://www.natureindex.com/news-blog/microsoft-academic-graph-discontinued-whats-next

https://opencitations.net

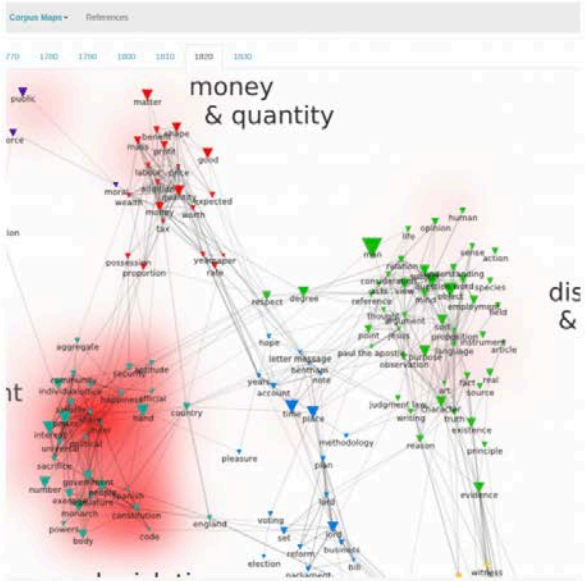
Des services logiciels orientés « fouille de textes »

CORTEXT

https://www.cortext.net



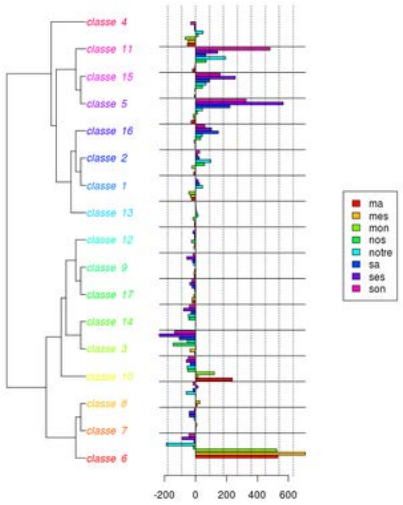
Hashtag cooccurrence network of tweets about Trump

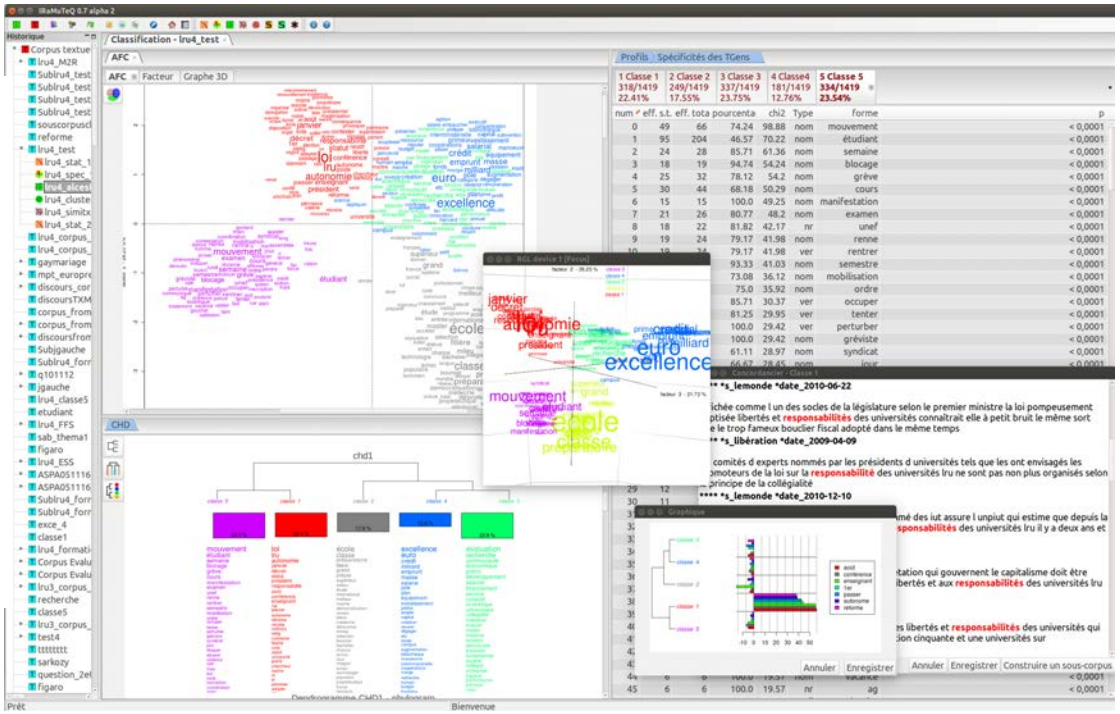


Mapping the Bentham Corpus

IRaMuTeQ

http://www.iramuteq.org





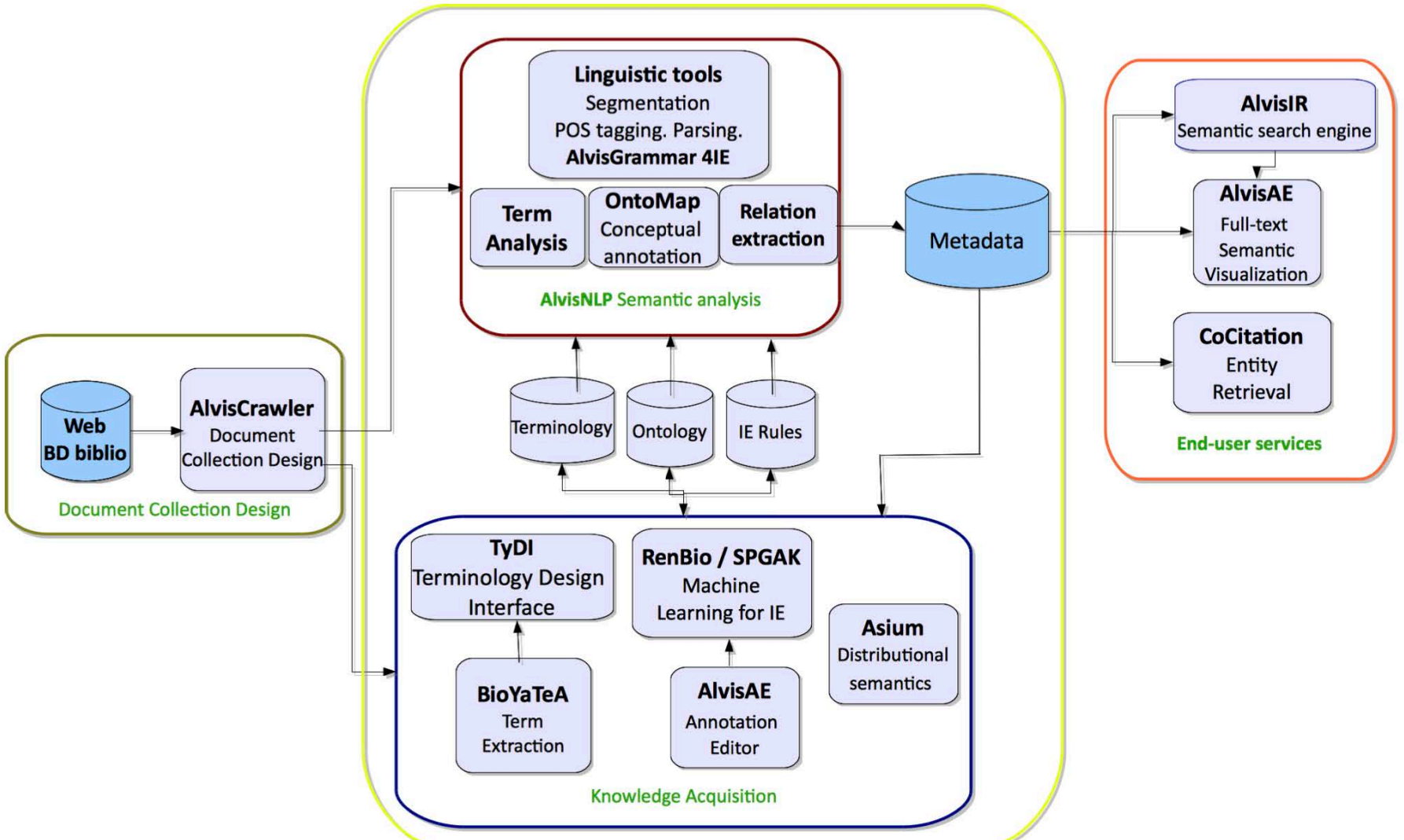
LERASS Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales

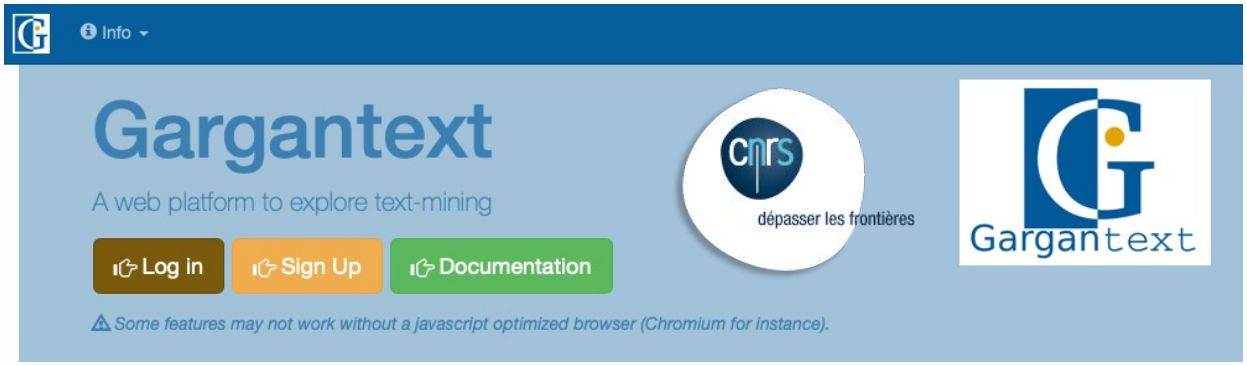
cnrs

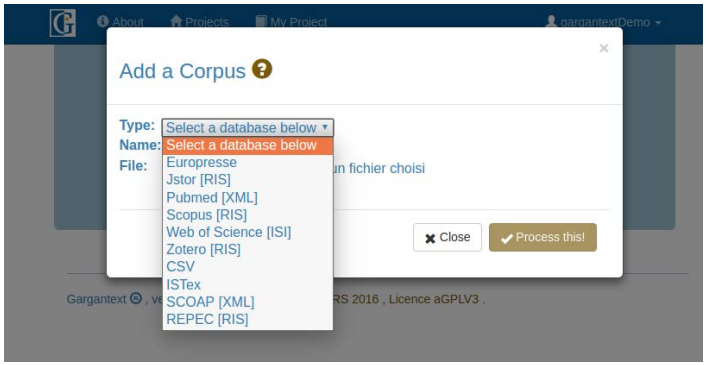
MaIAGE

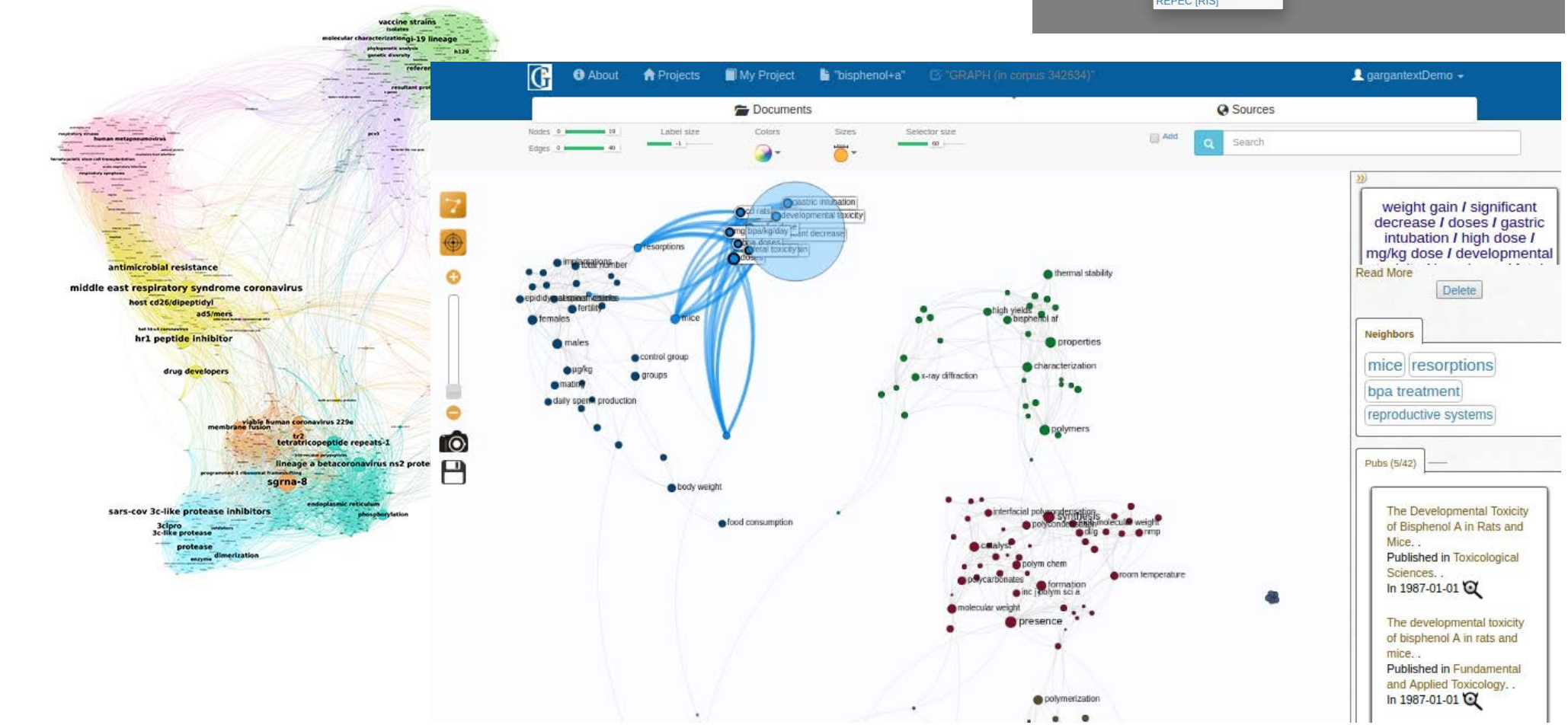
INRAE université PARIS-SACLAY

Plateforme Alvis



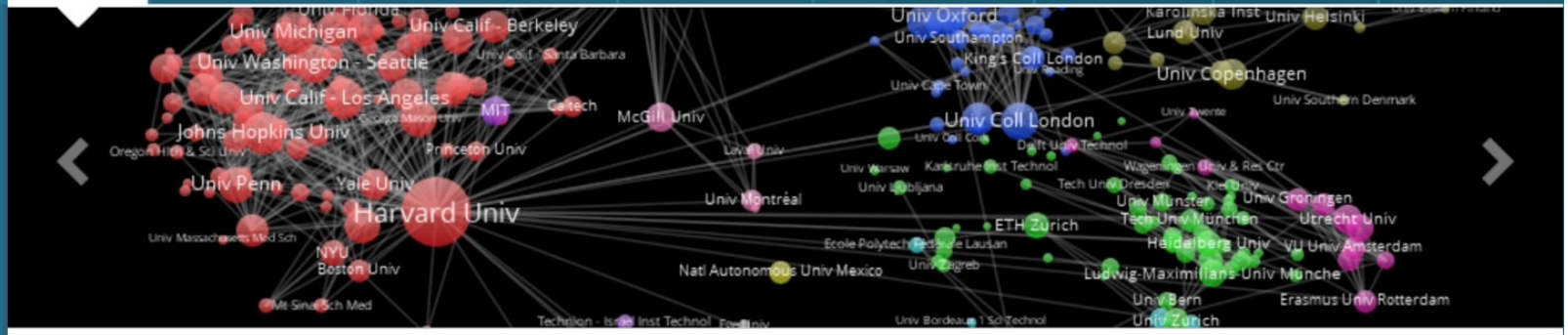






https://gargantext.org

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



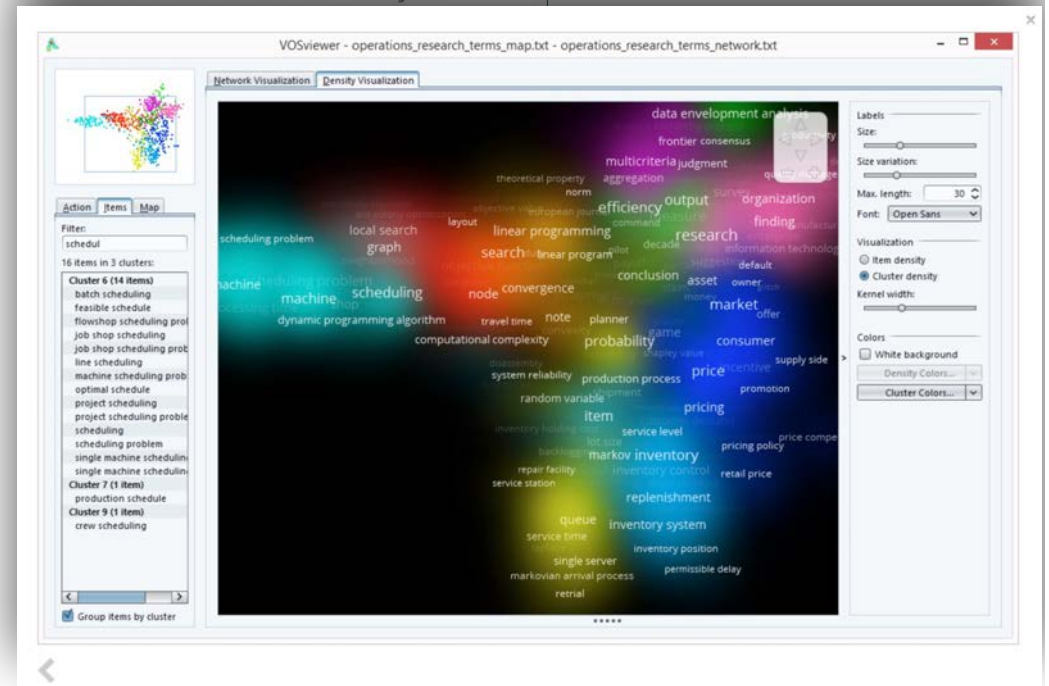
Welcome to VOSviewer

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.

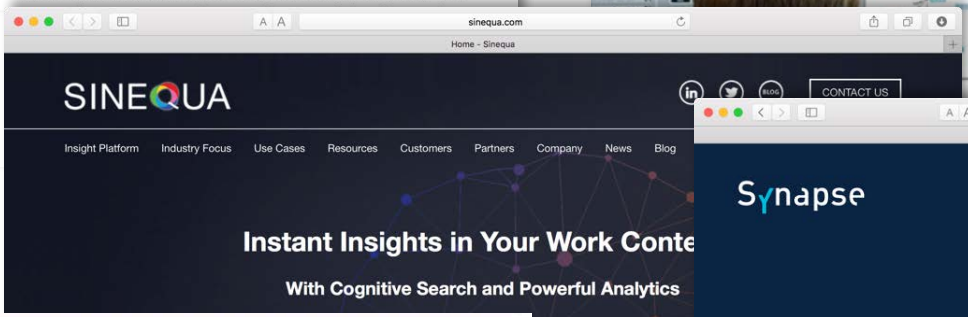
VOSviewer version 1.6.5

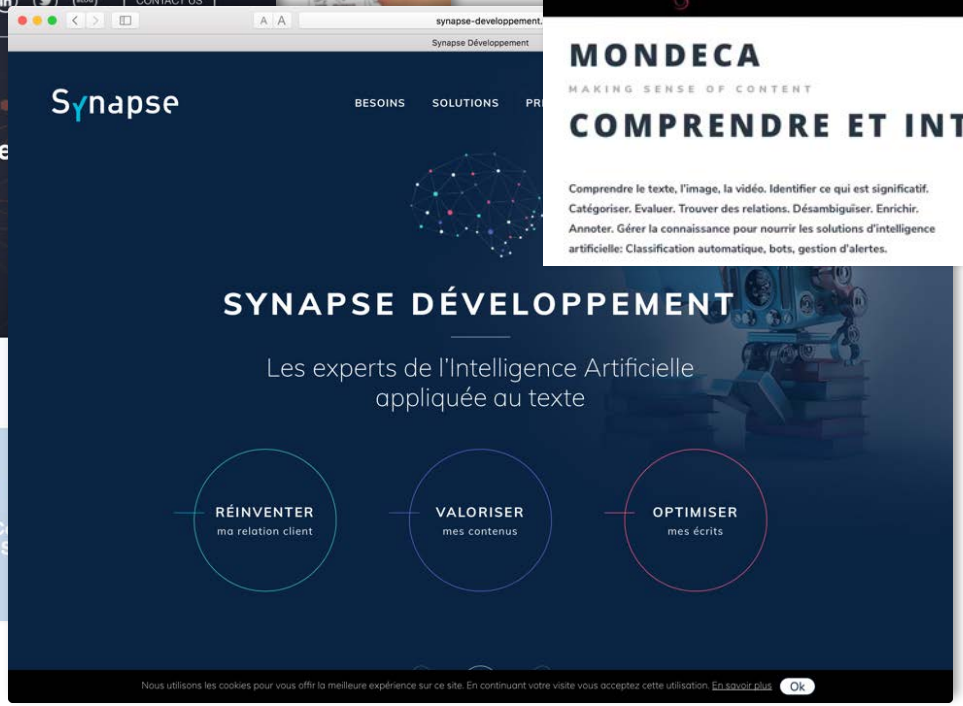
VOSviewer version 1.6.5 was released on September 28, 2016. Some of the improvements introduced in this version are listed below:

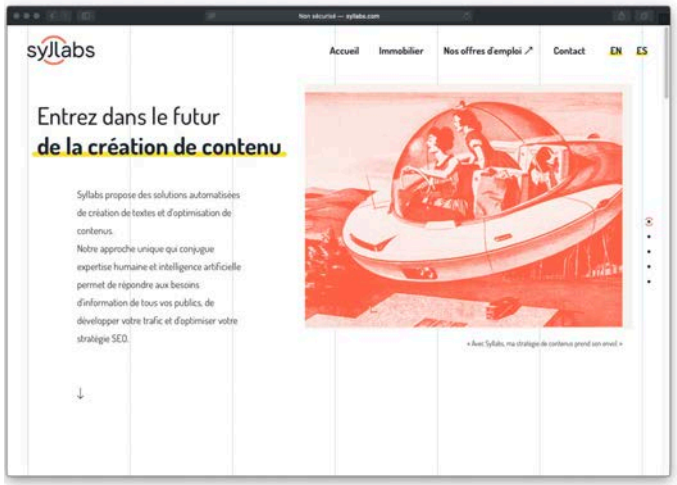
- **Overlay visualizations.** These popular visualizations have been made more prominently visible.
- **Maps based on bibliographic data.** Functionality for creating maps based on bibliographic data has been improved. Items can be filtered based on citation counts, and various types of overlay visualizations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



De nombreux acteurs industriels

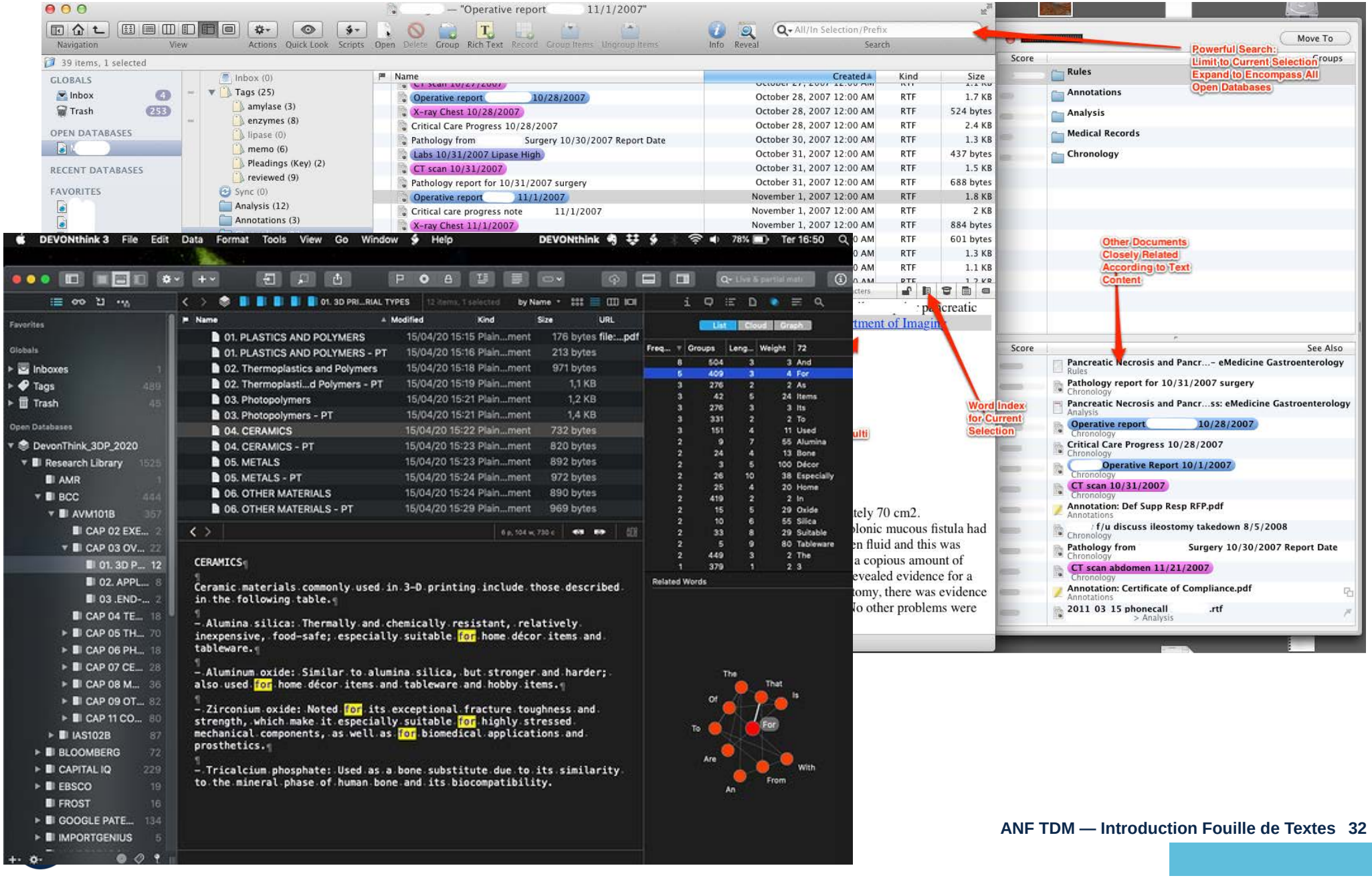






Devon Think

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

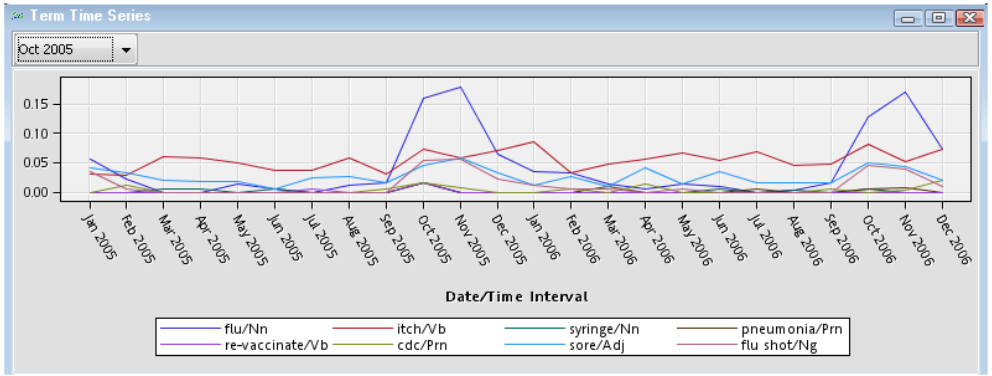


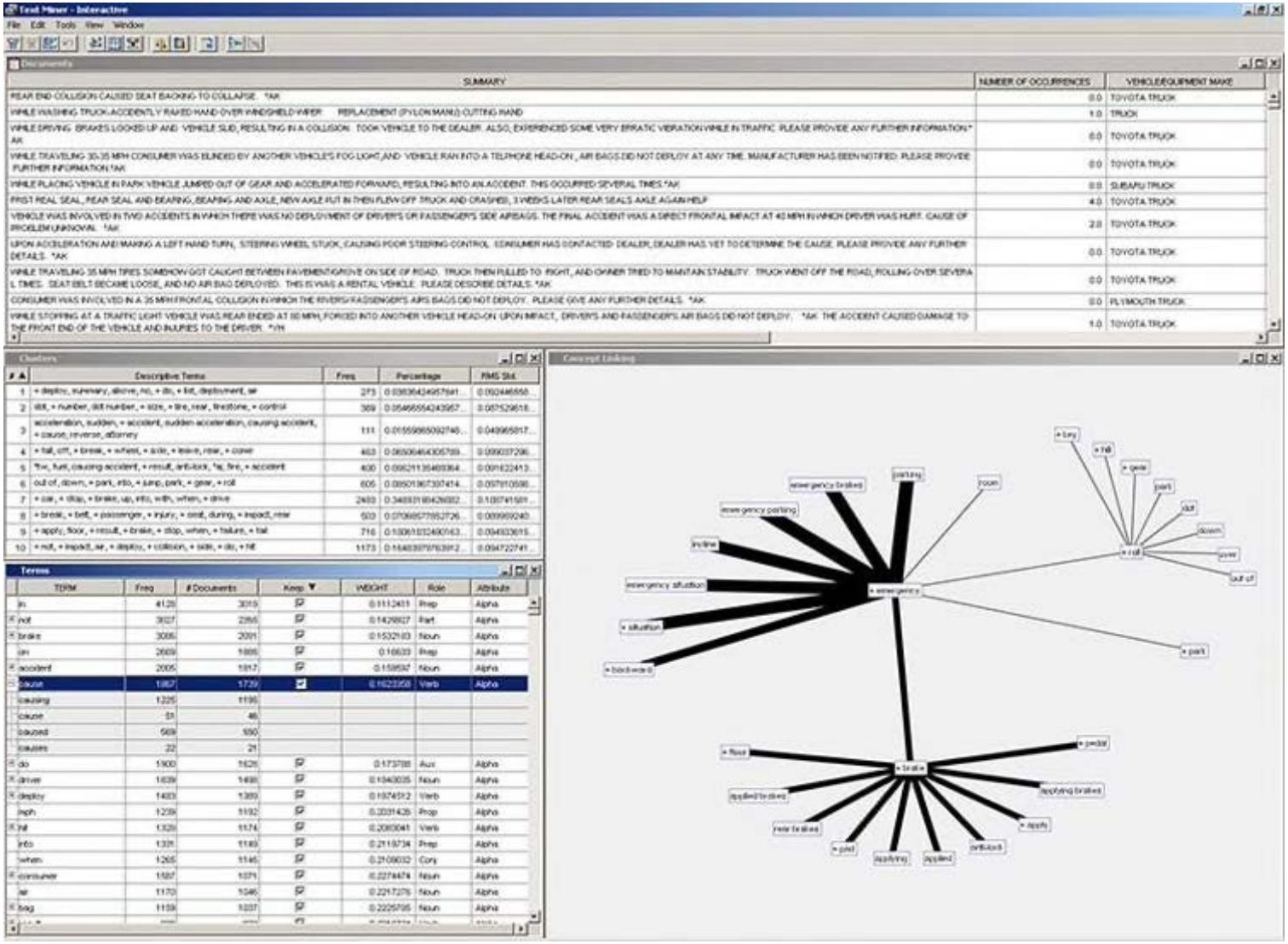
https://www.sas.com/fr_fr/software/text-miner.html

MENU

SAS® Text Miner

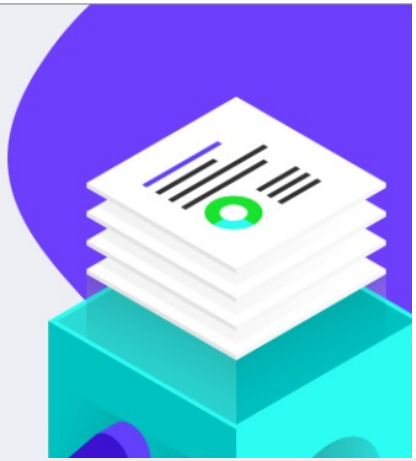
Mettez en évidence les informations dissimulées dans les données non structurées





cnrs

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

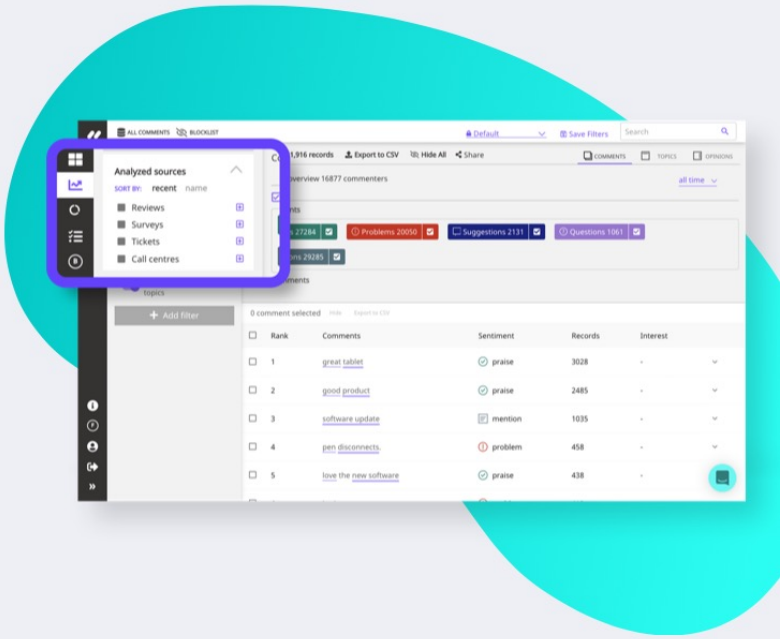
Instantly access the Praises, Problems, Suggestions, Questions, and Mentions in your feedback with accurate data categorization.

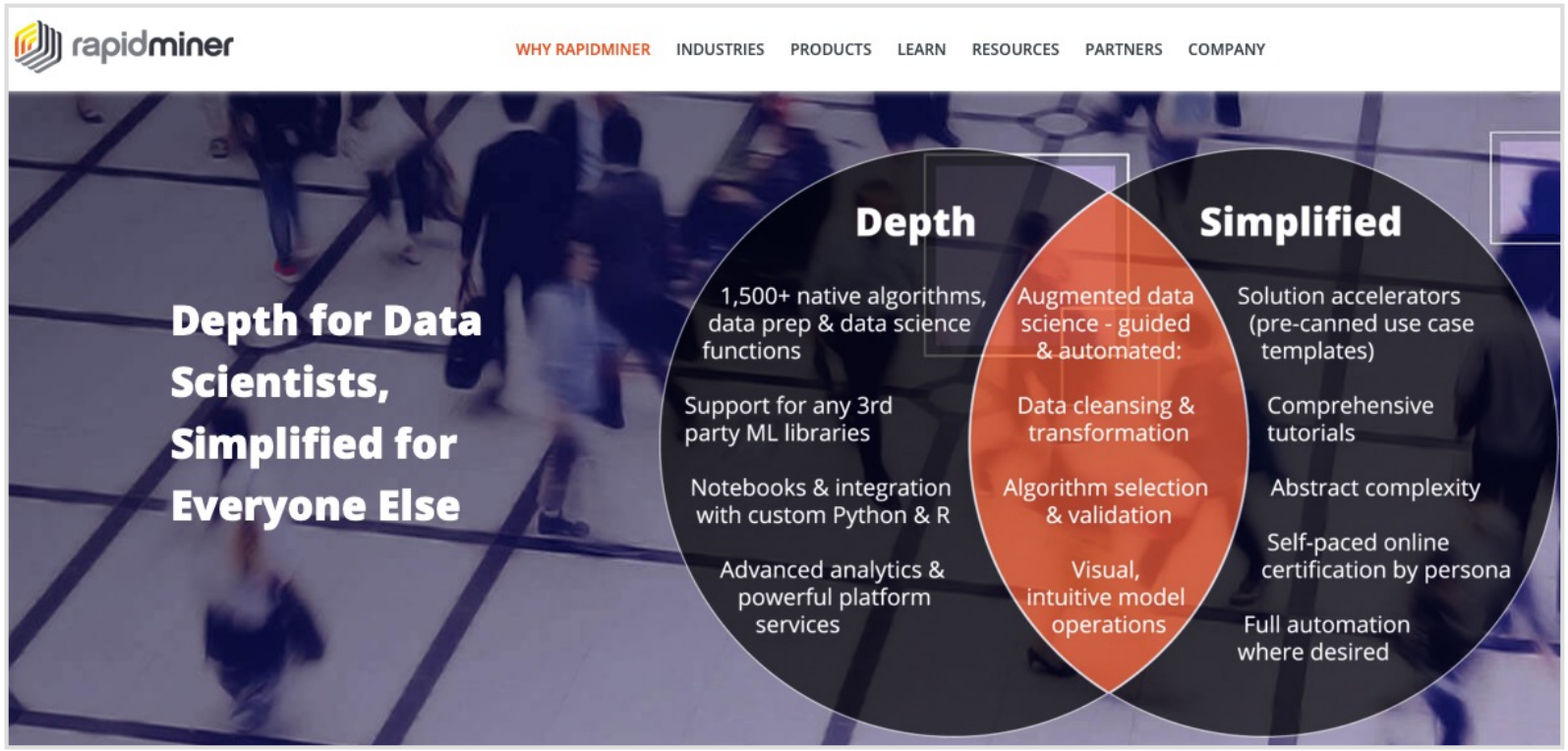
Multilingual processing

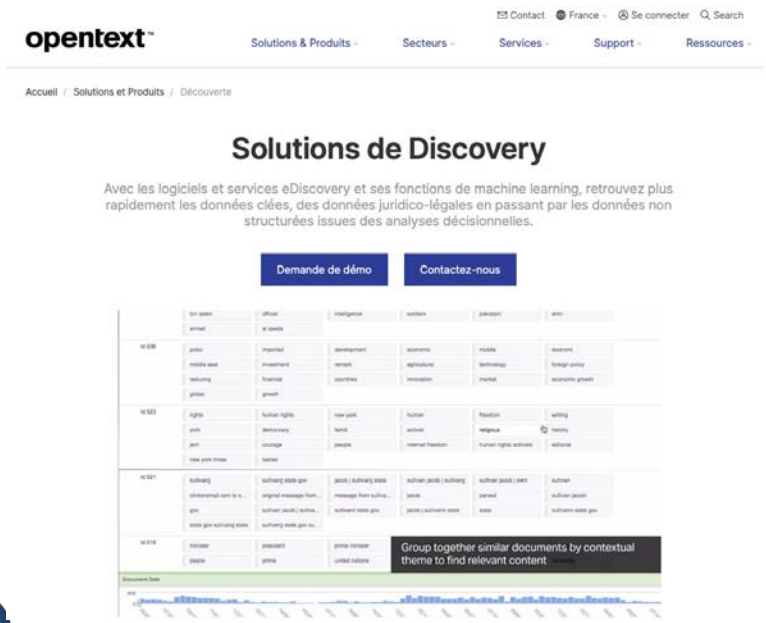
Analyze feedback with native proficiency in English and French, and top-of-the-line AI that translates 50+ other languages to English.

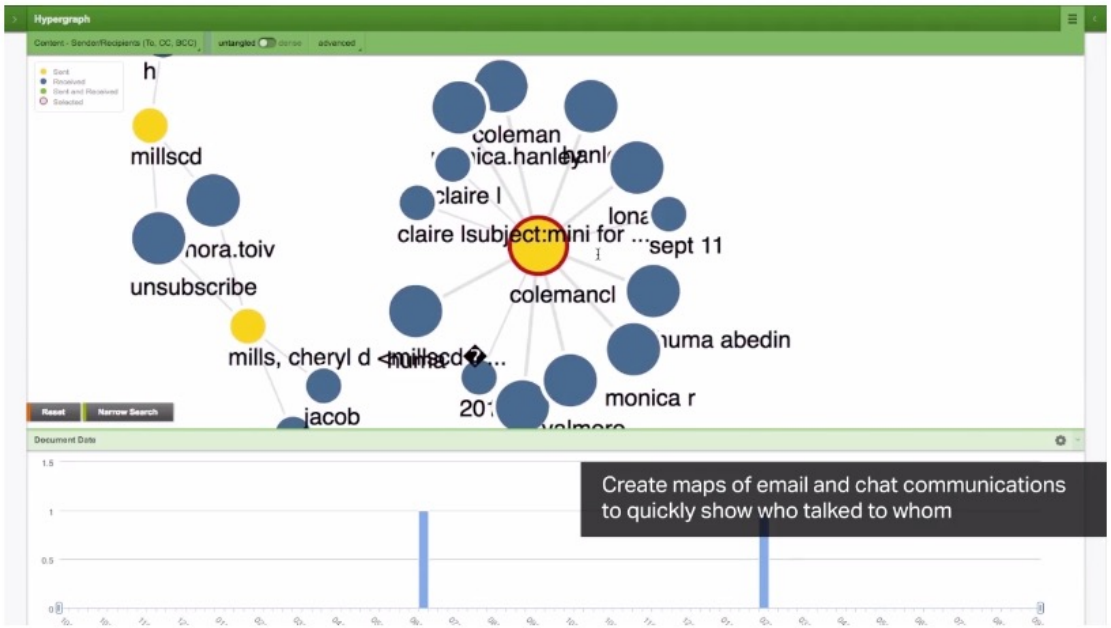
Advanced data filtering

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




rapidminer
WHY RAPIDMINER
INDUSTRIES
PRODUCTS
LEARN
RESOURCES
PARTNERS
COMPANY
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


opentext
Contact
France
Se connecter
Search
Solutions & Produits
Secteurs
Services
Support
Ressources
Accueil / Solutions et Produits / Découverte
Solutions de Discovery
Avec les logiciels et services eDiscovery et ses fonctions de machine learning, retrouvez plus rapidement les données clées, des données juridico-légales en passant par les données non structurées issues des analyses décisionnelles.
Demande de démo
Contactez-nous
Group together similar documents by contextual theme to find relevant content


Hypergraph
millscd
nora.toiv
unsubscribe
mills, cheryl d
jacob
coleman
claire l
claire lsubject:mini for
sept 11
colemancl
monica r
Reset
Narrow Search
Document Date
Create maps of email and chat communications to quickly show who talked to whom

Contact France Sign In Search

opentext™

Solutions & Produits Secteurs Services Support Ressources

Solutions de Discovery

Avec les logiciels et services eDiscovery et ses fonctions de machine learning, retrouvez plus rapidement les données clées, 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

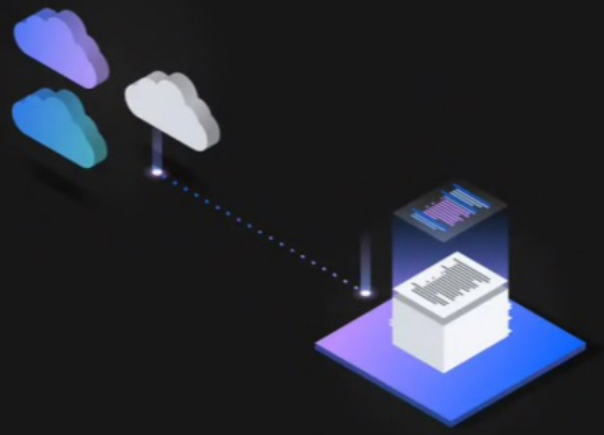
https://www.ibm.com/cloud/watson-natural-language-understanding

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.

Overview

Powerful Insight Extraction

Get underneath the topics mentioned in your data by using text analysis to extract keywords, concepts, categories and more.

Learn more

Extensive Language Support

Analyze your unstructured data in more than thirteen languages.

Learn more

High-Accuracy Extraction

Out-of-the-box machine learning models for text mining provide a high degree of accuracy across your content.

Learn more

Why Watson NLU?

Deploy Anywhere

Deploy Watson Natural Language Understanding behind your firewall or on any cloud.

Learn more

Domain Customization

Train Watson to understand the language of your business and extract customized insights with Watson Knowledge Studio.

Learn more

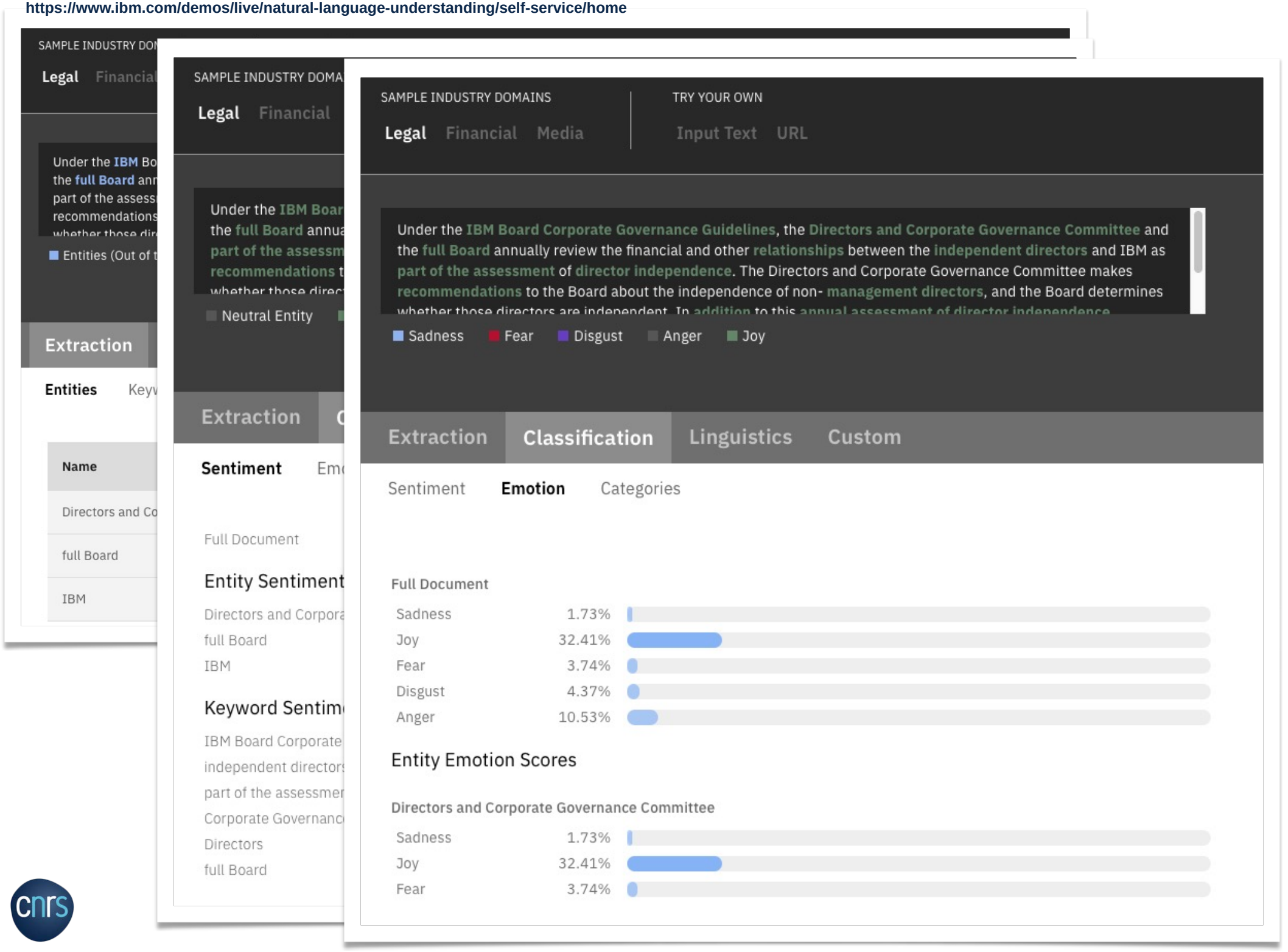
Data Control

Maintain ownership of your data with the assurance that your data is safe and secure. IBM will not collect or store your data.

Learn more

Let's talk

https://www.ibm.com/demos/live/natural-language-understanding/self-service/home



IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 14, NO. 6, NOVEMBER/DECEMBER 2008

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.

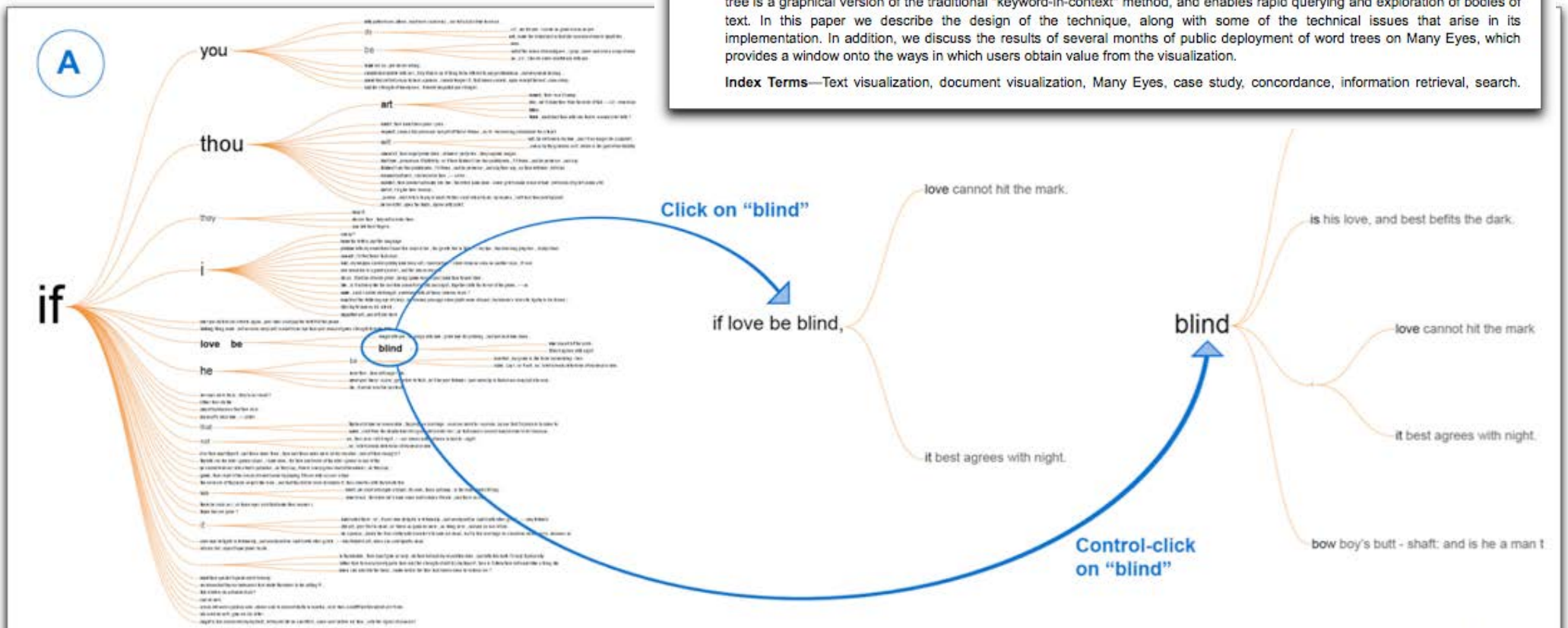
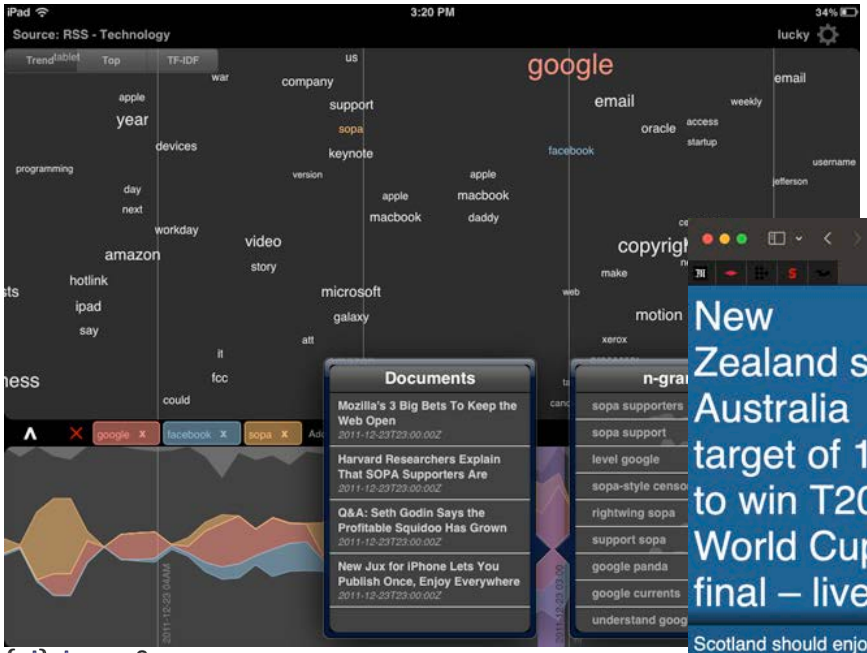
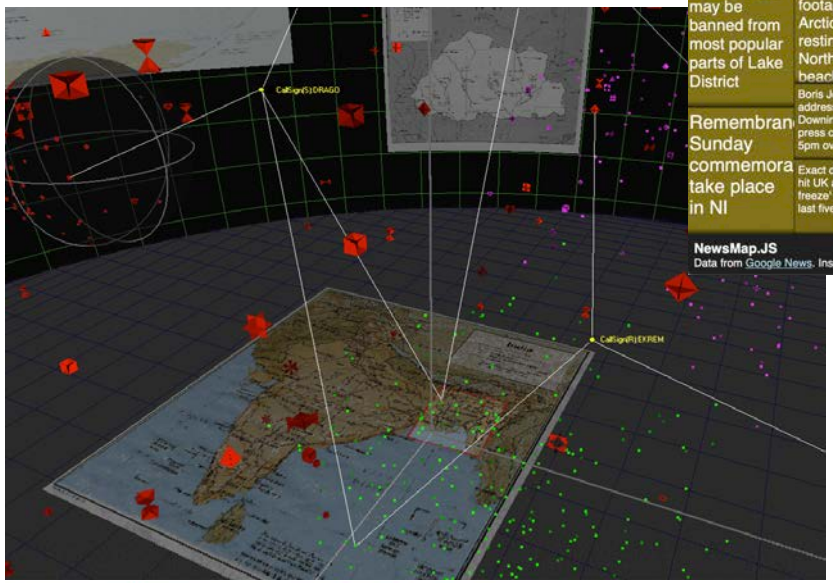


Fig 3. Sequence showing some of the interaction options in the word tree. In figure A, the user has typed the word "if" in *Romeo and Juliet*. In B, the user has clicked on "blind," which appears in one of the branches under "if." This causes the visualization to recenter to the longer phrase "if love 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.

http://www-958.ibm.com/software/data/cognos/manyeyes/page/Word_Tree.html



LuckySort



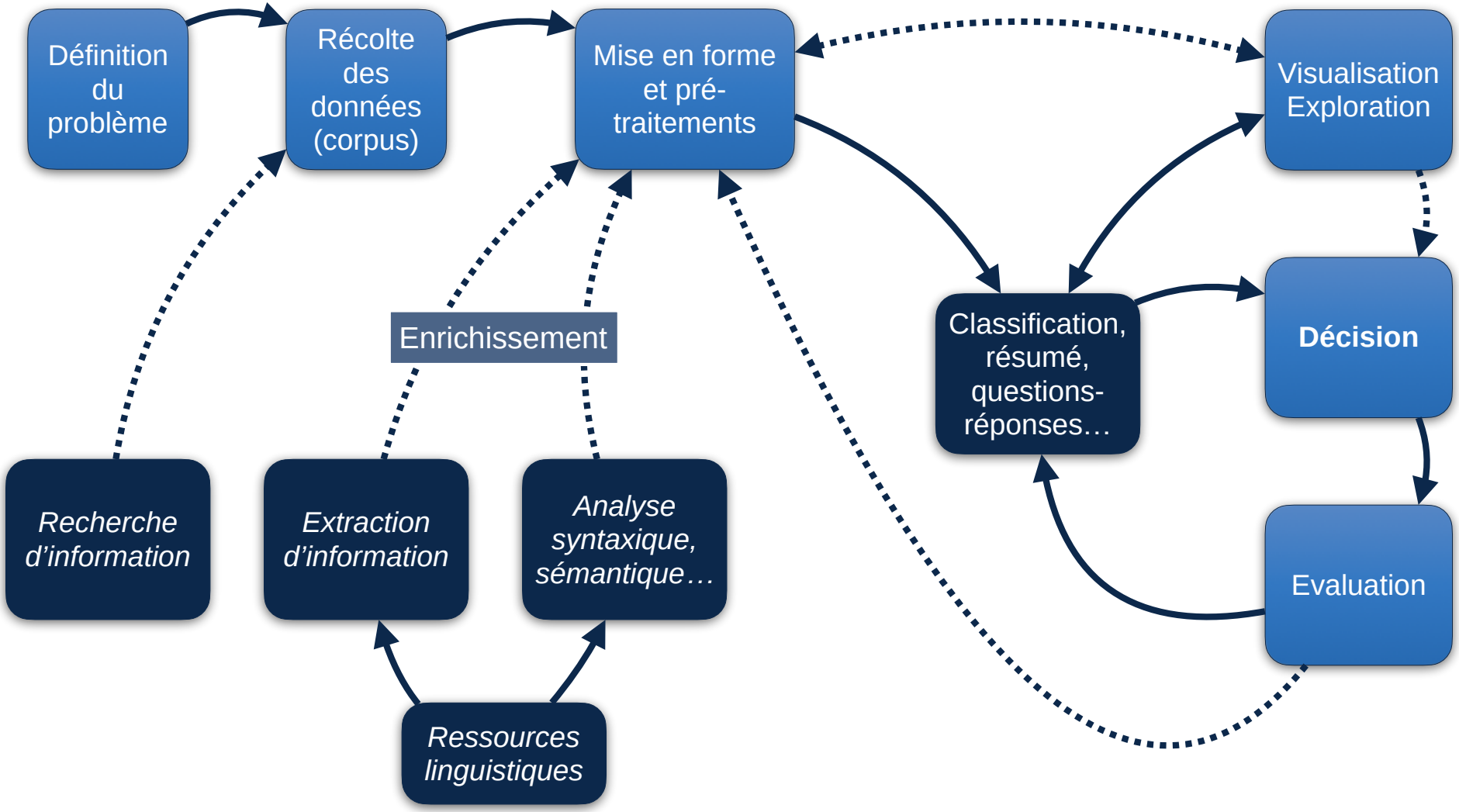
STARLIGHT
Visual Information System™

newsmap

https://newsmap-js.herokuapp.com

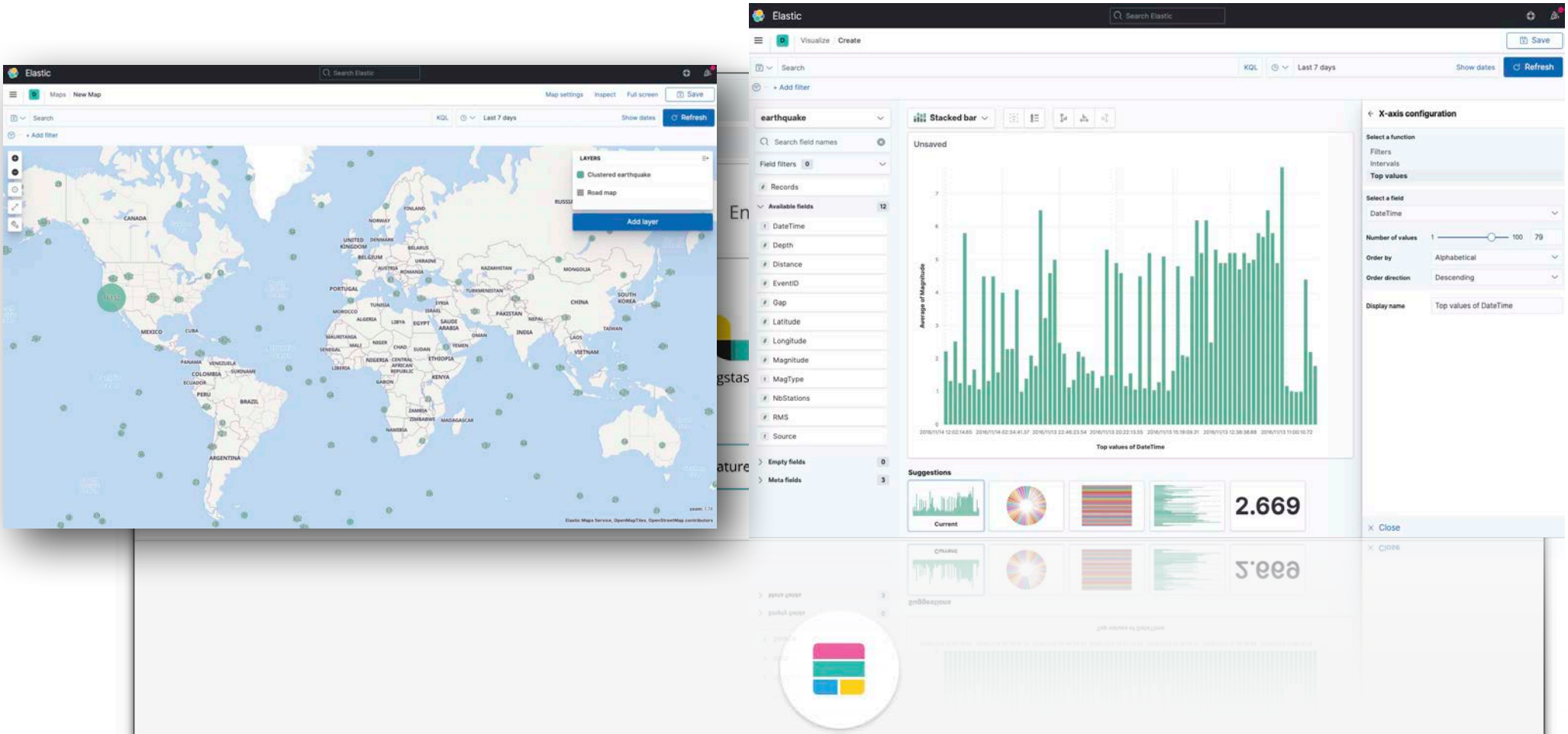
Des environnements logiciels pour le développement et l'expérimentation

Un processus de fouille de textes



Indexation et recherche d'information

https://www.elastic.co



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.

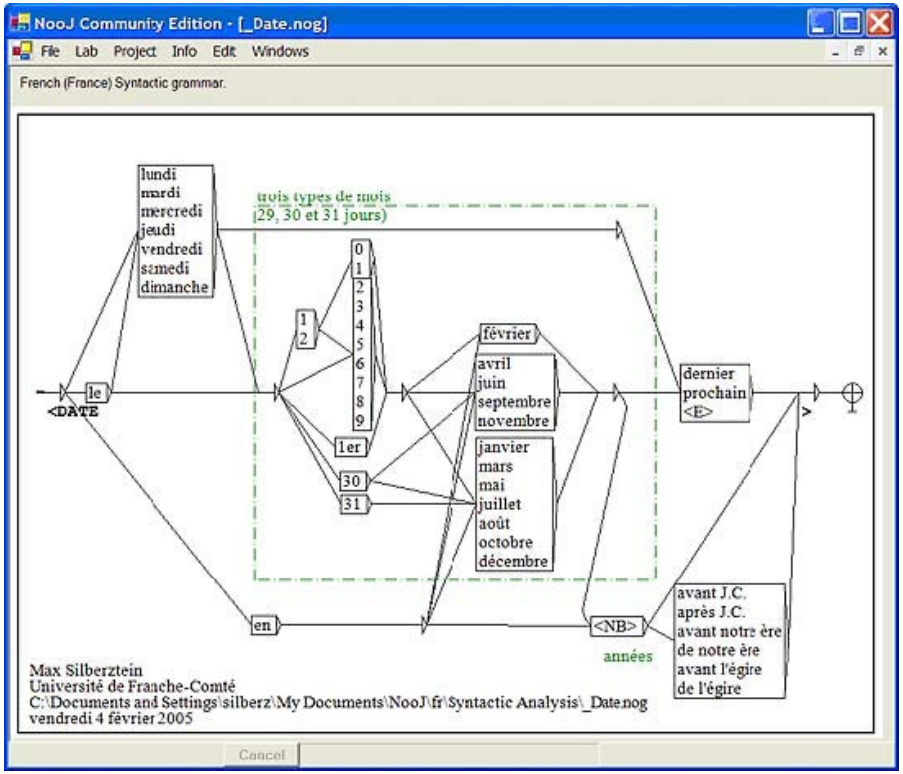
s'appuie sur Apache Lucene

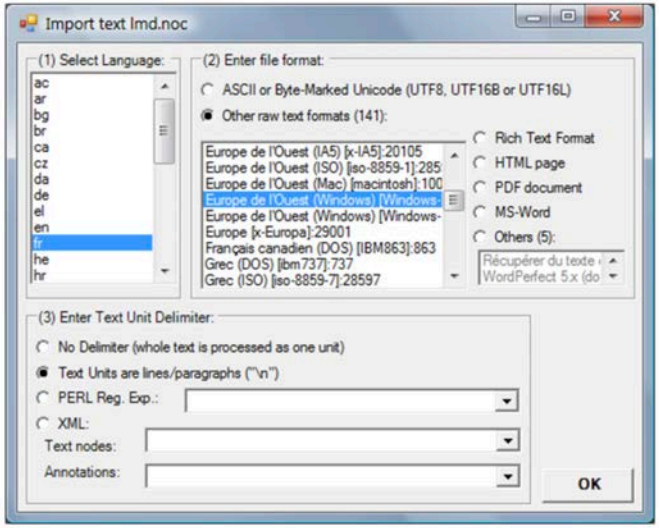
https://dl.istex.fr

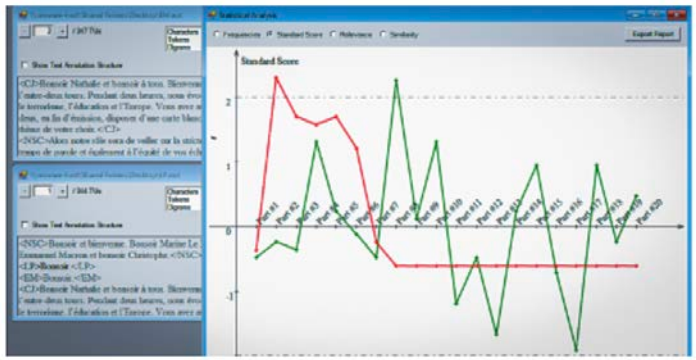
Traitement de corpus avec grammaires et dictionnaires

NooJ

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







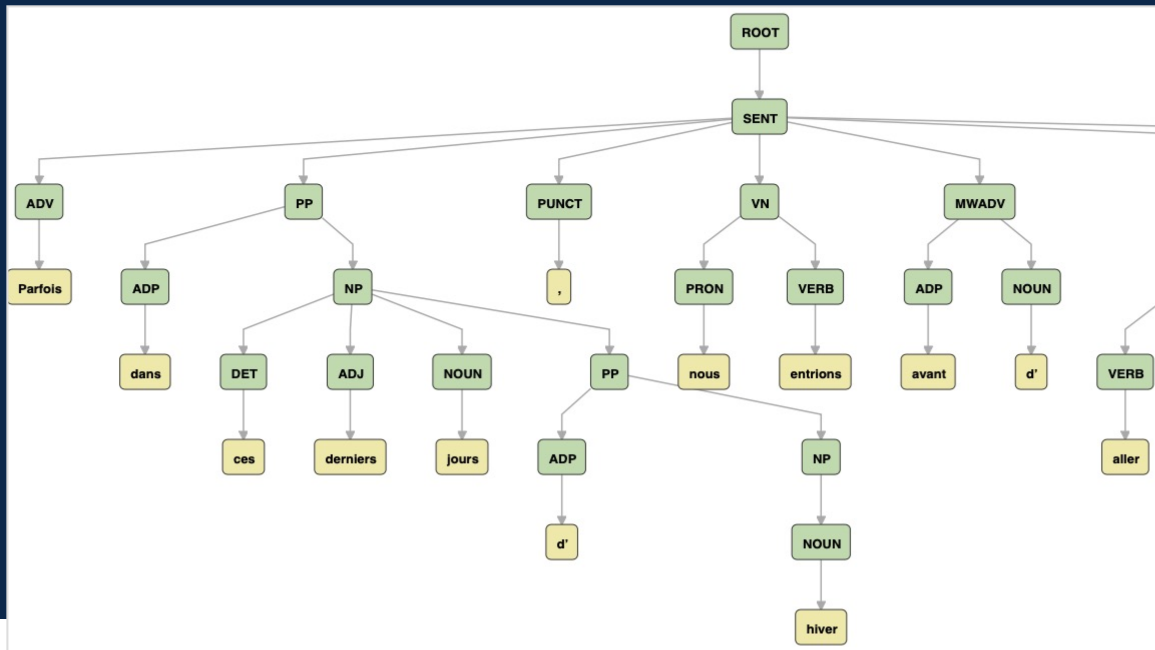
https://www.nooj-association.org

Enrichissement et annotations « linguistiques »

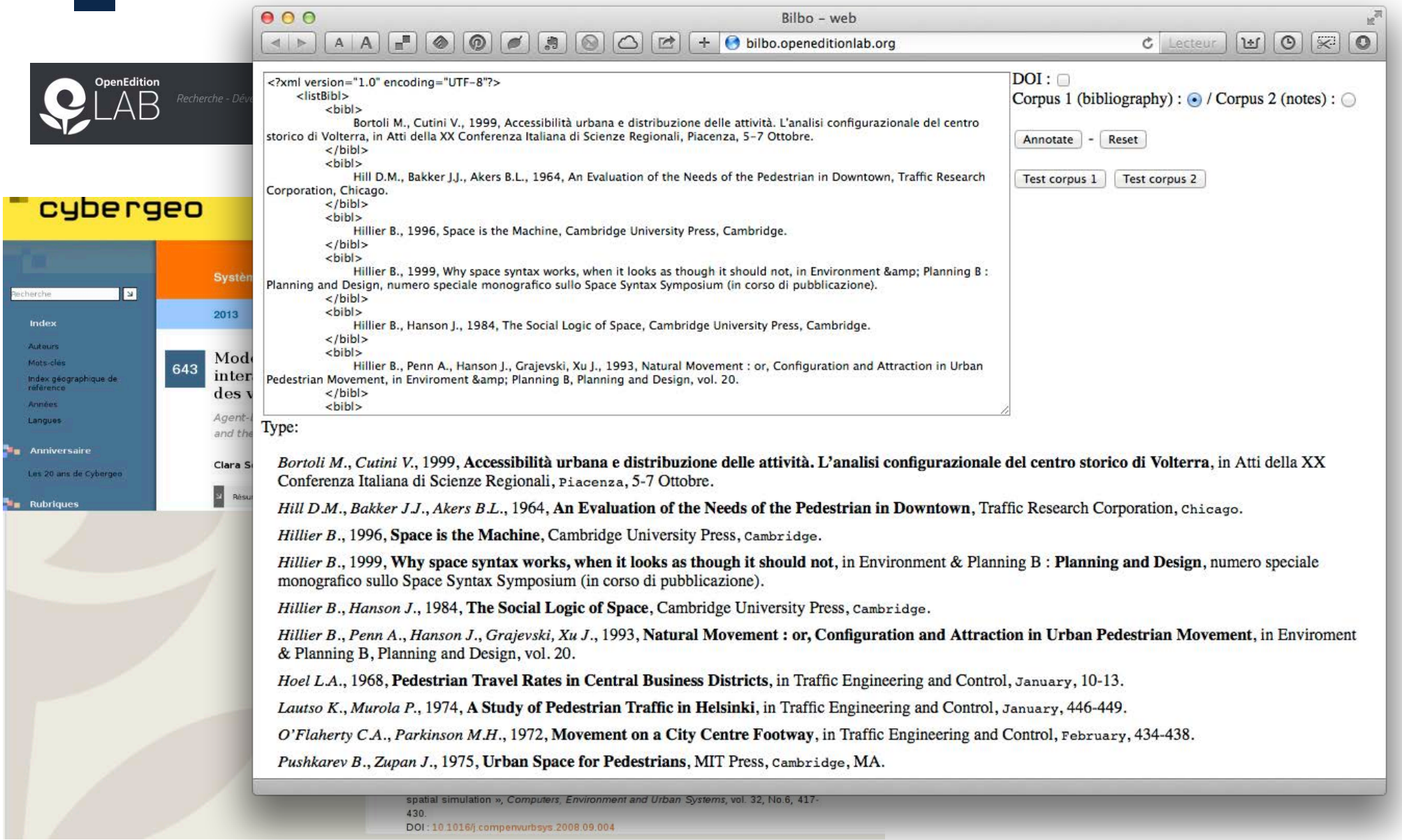
Dépendances

Unités lexicales et parties du discours

Constituants



Structuration automatique de références bibliographiques

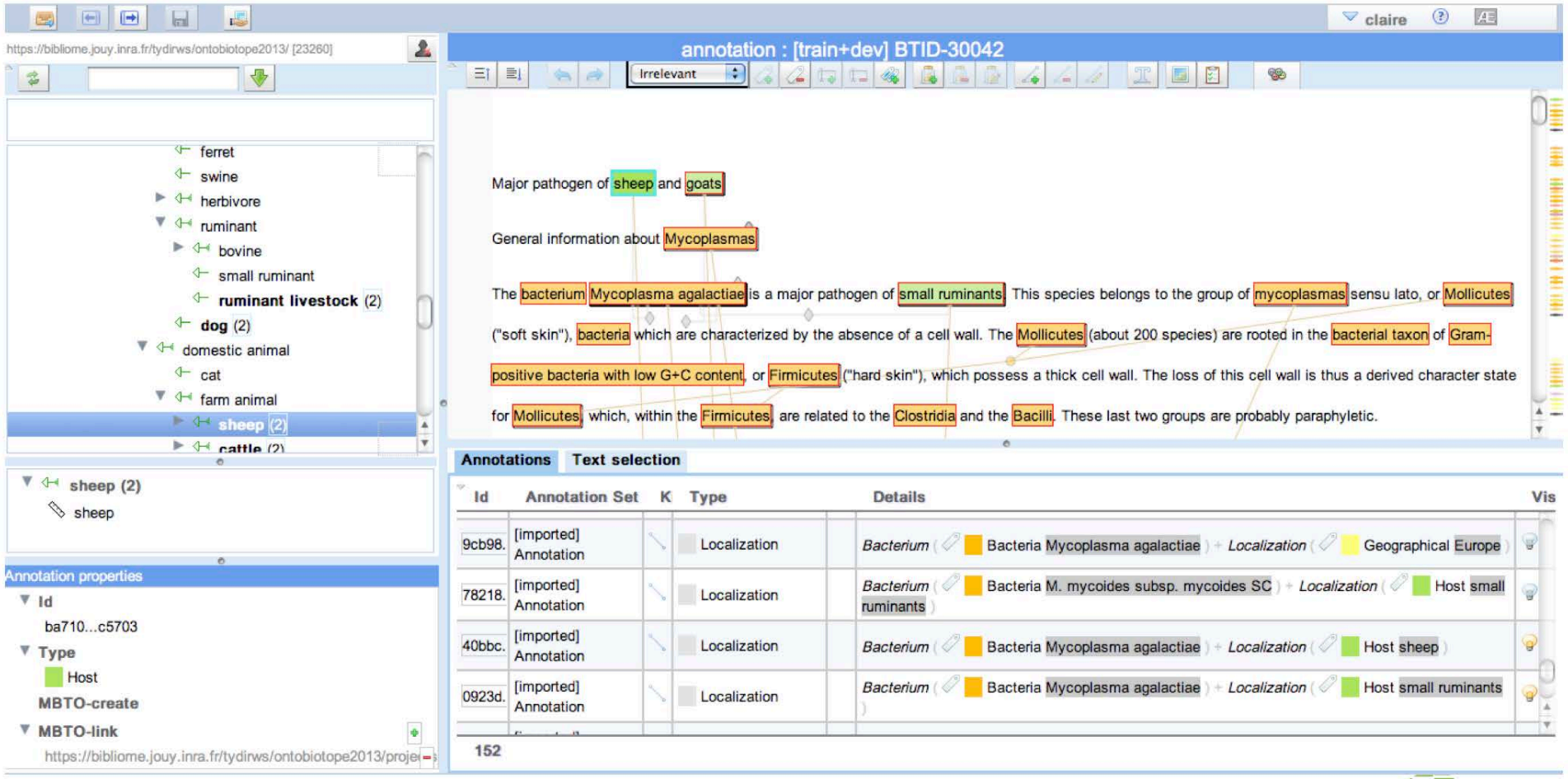


Test : http://bilbo.openeditionlab.org

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

AlvisAE (1)

Annotation manuelle collective de textes.



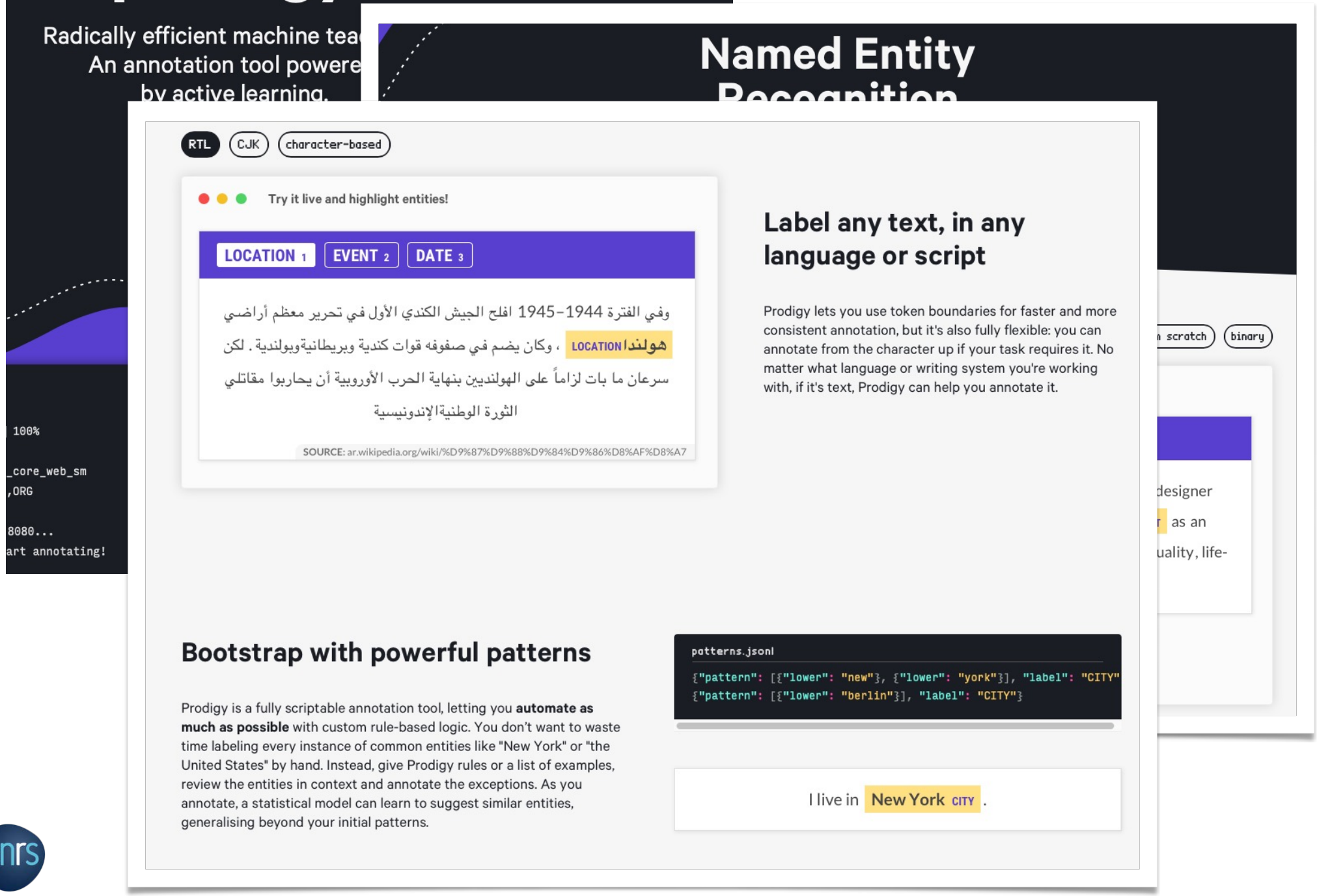
INRA SCIENCE & IMPACT

cnrs

prodigy

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

https://prodi.gy/



Classification

https://www.cs.waikato.ac.nz/ml/weka/

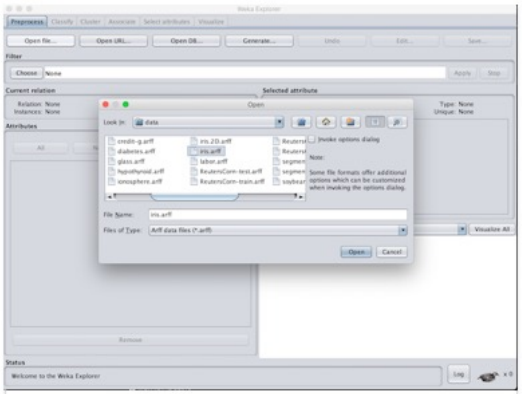
WEKA

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
plethora

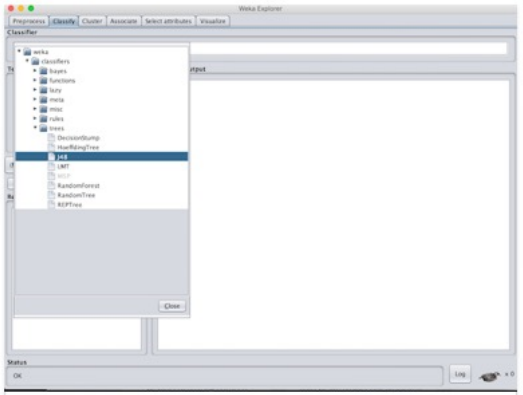
Machine Learning without Programming

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



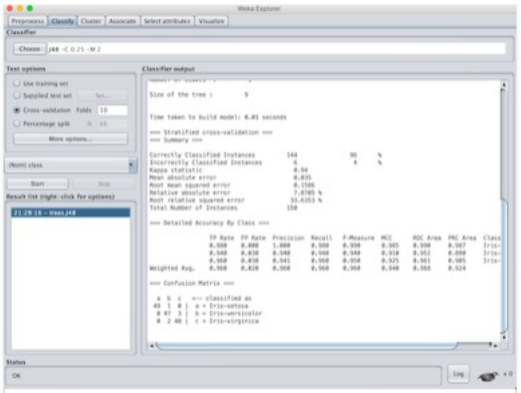
Open a dataset

First, we open the dataset that we would like to evaluate.



Choose a classifier

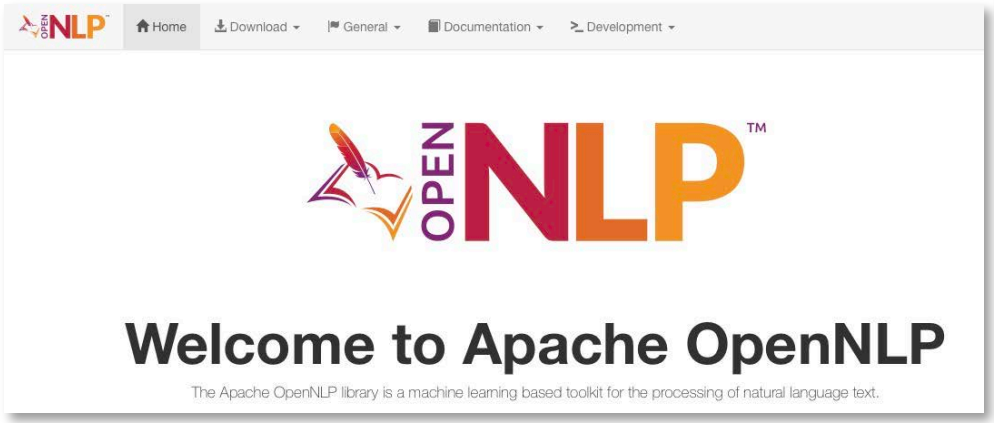
Second, we select a learning algorithm to use, e.g., the J48 classifier, which learns decision trees.

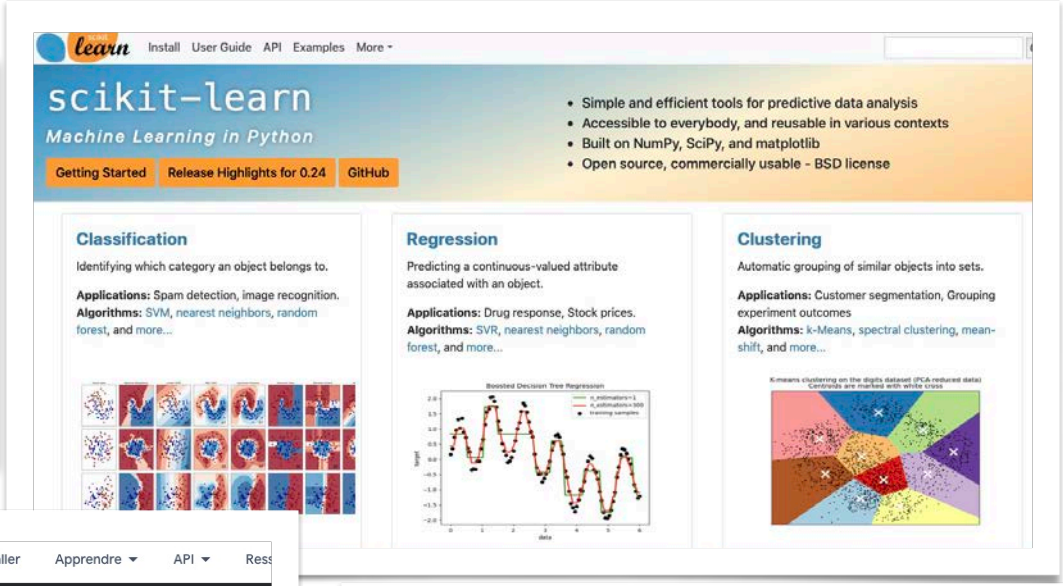


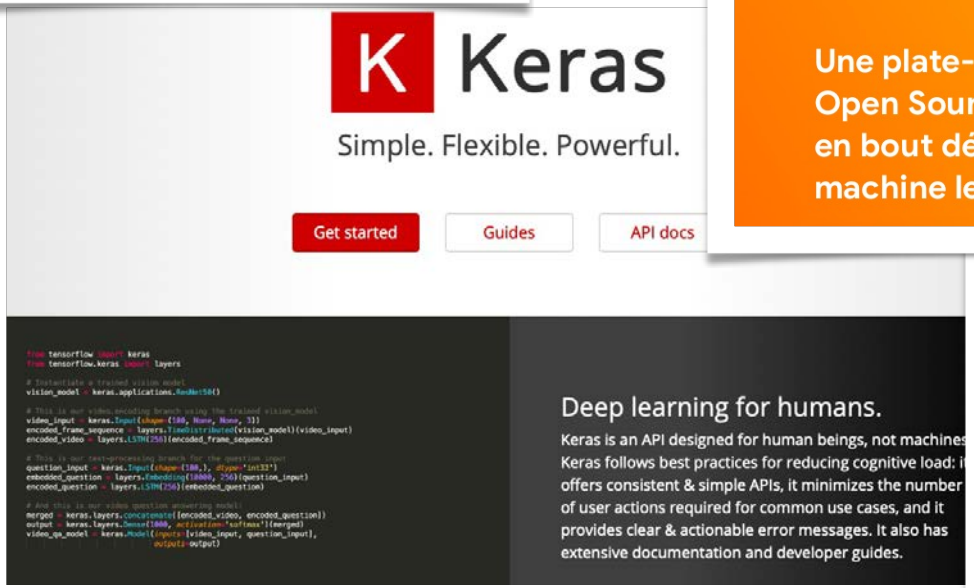
Evaluate predictive accuracy

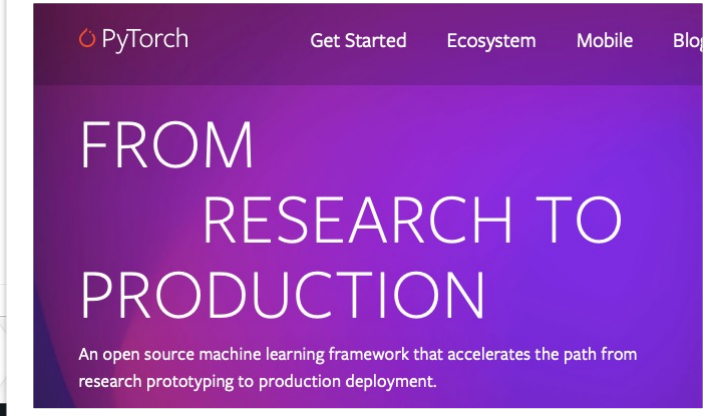
Finally, we run a 10-fold cross-validation evaluation and obtain an estimate of predictive performance.

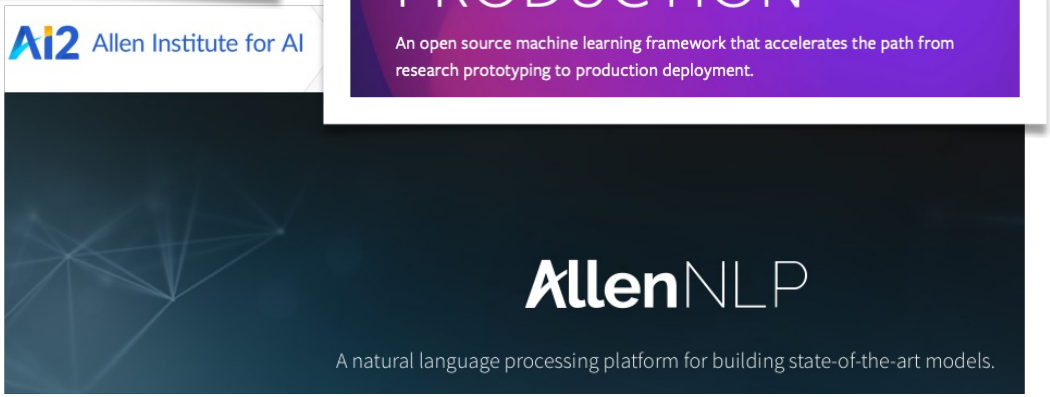
Des bibliothèques Python (ou Java, C++, Swift)

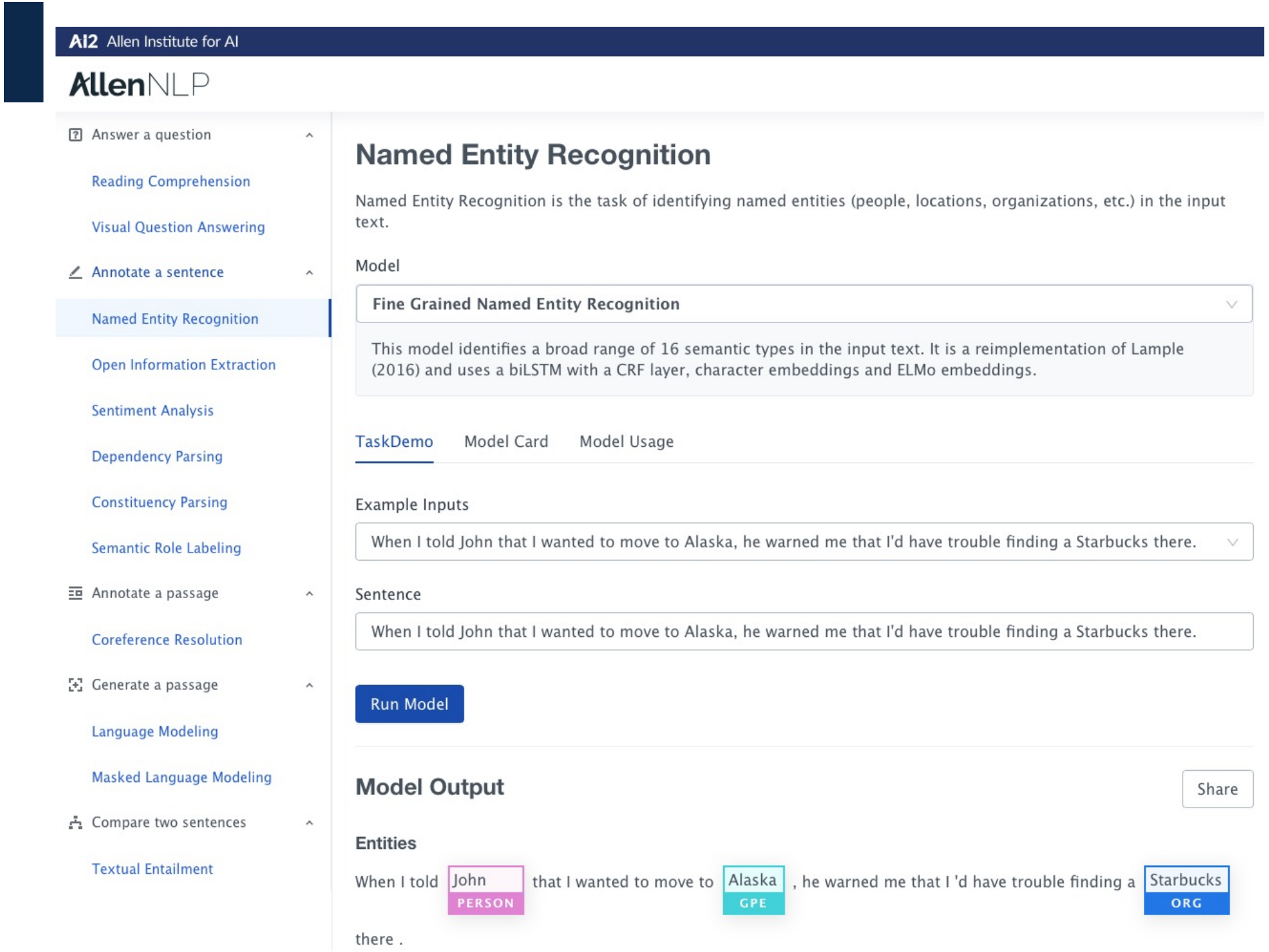











AI2 Allen Institute for AI
AllenNLP
Answer a question
Reading Comprehension
Visual Question Answering
Annotate a sentence
Named Entity Recognition
Open Information Extraction
Sentiment Analysis
Dependency Parsing
Constituency Parsing
Semantic Role Labeling
Annotate a passage
Coreference Resolution
Generate 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 reimplementation 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 PERSON that I wanted to move to Alaska GPE , he warned me that I 'd have trouble finding a Starbucks ORG there .

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

The doctor ran to the emergency room to see the

Sentence

The doctor ran to the emergency room to see the

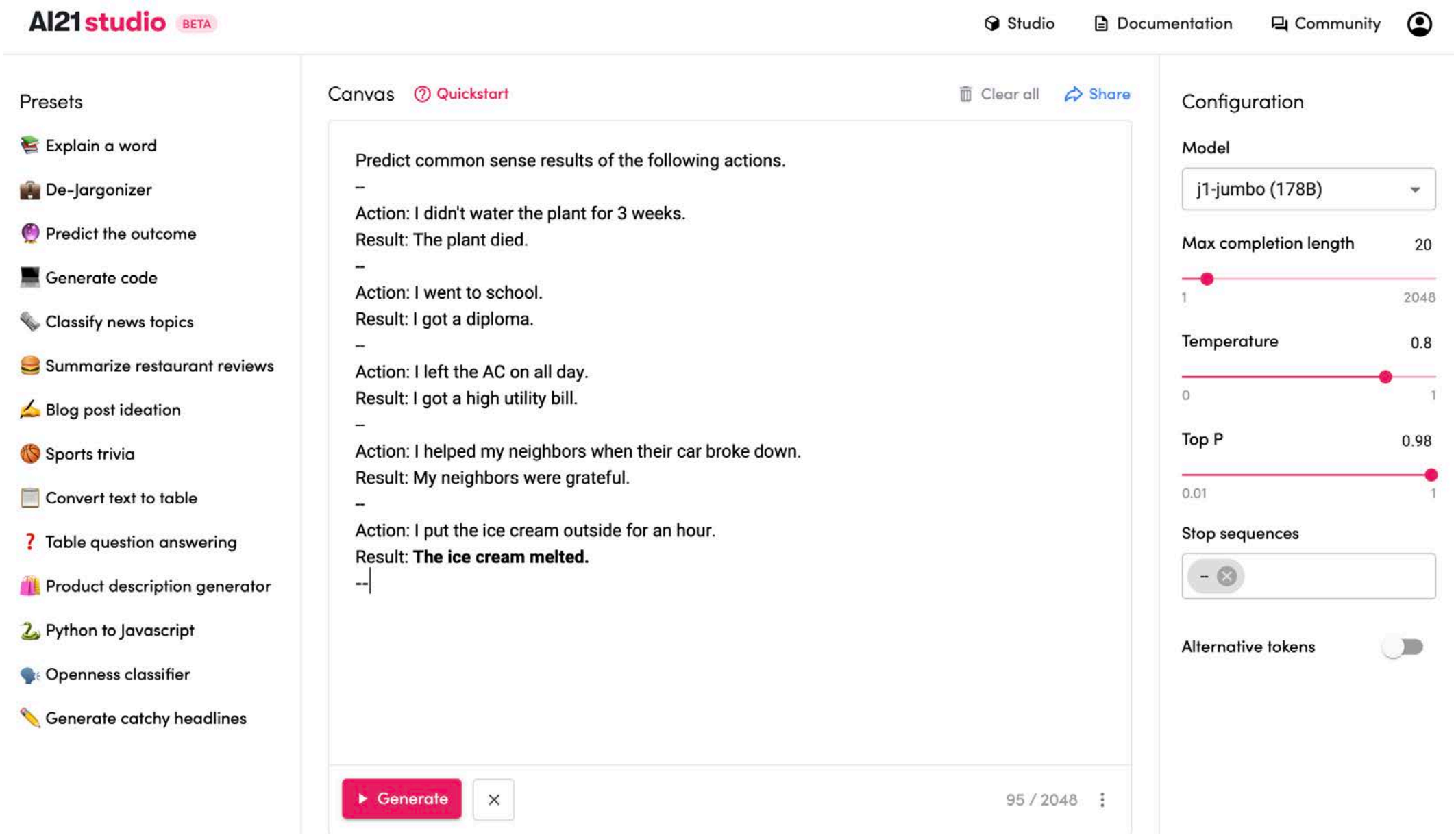
Run Model

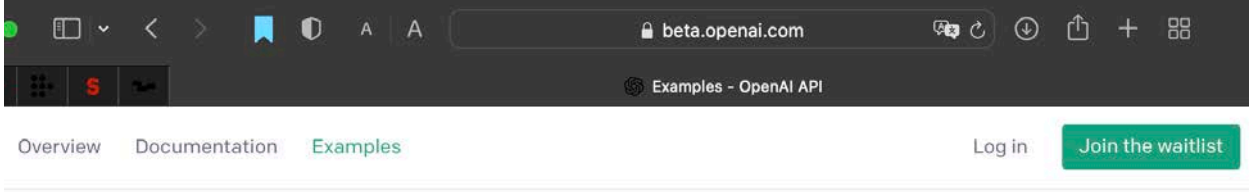
Model Output

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 %

https://studio.ai21.com/playground





OpenAI
@OpenAI

https://openai.com/

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 u

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.

cnrs

https://openai.com/api/

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 Fastest	Babbage	Curie	Davinci Most powerful
$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

http://www.nltk.org

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.”

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

Please consult the README file included with each corpus for further information.

Unicode Version 7.0.0 character properties in Perl [download | source]
100266; author: ; copyright: ; license: ;

ligner (Sultan et al. 2015) subset of the Paraphrase Database. [download | source]
711; author: ; copyright: ; license: Creative Commons Attribution 3.0 Unported (CC-BY);

[download | source]
3; author: Jan Strunk; copyright: ; license: ;

dor de Sufixos da Lingua Portuguesa) [download | source]
hor: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christian Huyck; copyright: ; license: ;

es [download | source]
00510; author: ; copyright: ; license: ;

ad | source]
: 6785405; author: ; copyright: ; license: ;

nker (Maximum entropy) [download | source]
; size: 13404747; author: ; copyright: ; license: ;

download | source]
10961490; author: ; copyright: ; license: ;

del [download | source]
size: 24516205; author: ; copyright: ; license: ;

nload | source]
ze: 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: ;

NLTK

http://www.nltk.org

Some simple things you can do with NLTK

Tokenize and tag some text:

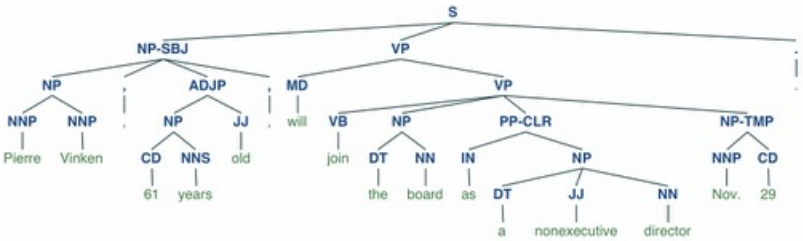
```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)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



NLTK

```
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)
```

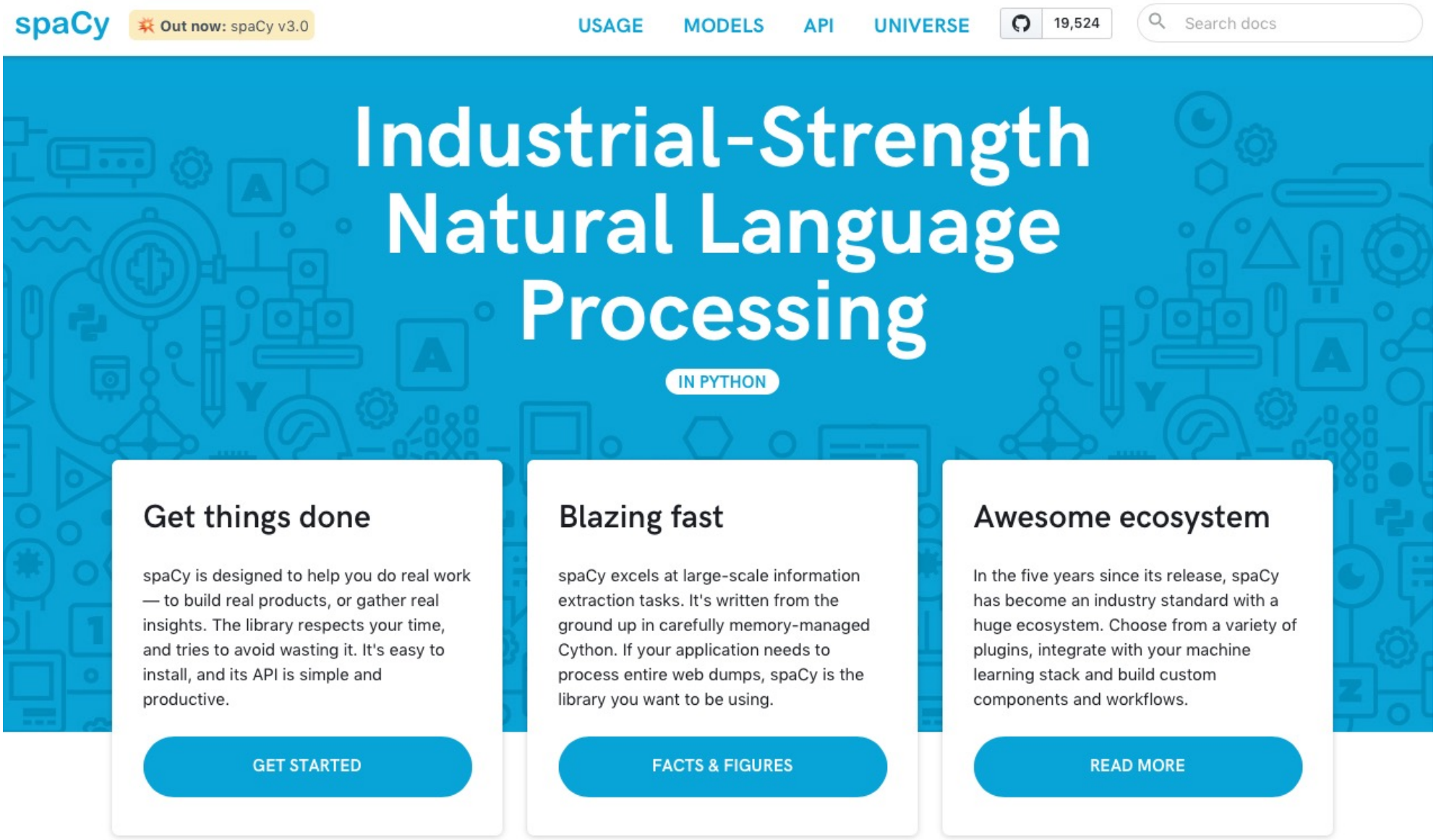
We can now train our classifier on the training set, and subsequently output the evaluation results:

```
>>> 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

Spacy

https://spacy.io/



Spacy

https://spacy.io/

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

Fonctionnalités de Spacy

https://spacy.io/

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.

tm - Text Mining Package

http://tm.r-forge.r-project.org/

tm (shorthand for *Text Mining Infrastructure in R*) provides a framework for text mining applications within R.

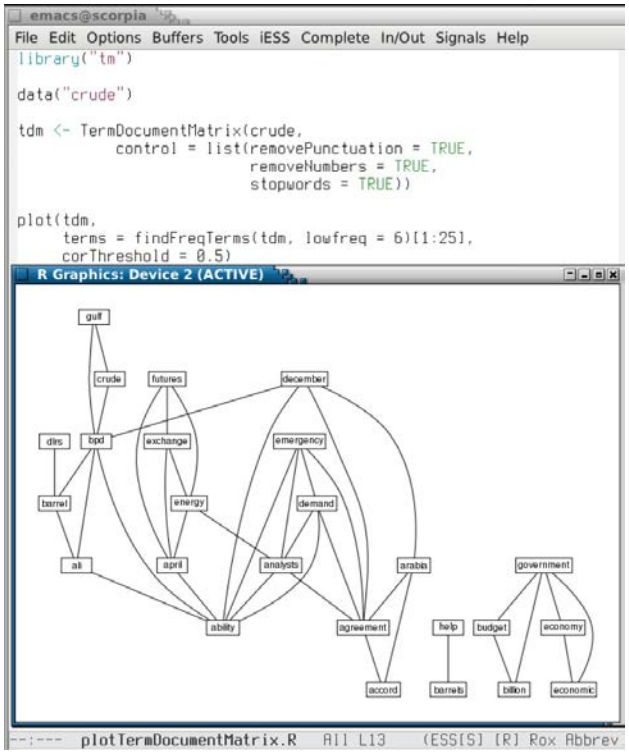
The tm package offers functionality for managing text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats in R. The package has integrated database back-end support to minimize memory demands. An advanced meta data management is implemented for collections of text documents to alleviate the usage of large and with meta data enriched document sets.

The package provides native support for reading in several classic file formats (e.g. plain text, PDFs, or XML files). There is also a plug-in mechanism to handle additional file formats.

The data structures and algorithms can be extended to fit custom demands, since the package is designed in a modular way to enable easy integration of new file formats, readers, transformations and filter operations.

tm provides easy access to preprocessing and manipulation mechanisms such as whitespace removal, stemming, or stopword deletion. Further a generic filter architecture is available in order to filter documents for certain criteria, or perform full text search. The package supports the export from document collections to term-document matrices.

tm is freely available under the GNU General Public License (GPL).



openNLP: Apache OpenNLP Tools Interface

An interface to the Apache OpenNLP tools (version 1.5.3). The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text written in Java. common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. See <https://opennlp.a information.

Version:	0.2-7
Imports:	NLP (≥ 0.1-6.3), openNLPdata (≥ 1.5.3-1), rJava (≥ 0.6-3)
Suggests:	openNLPmodels.en
Published:	2019-10-26
Author:	Kurt Hornik [aut, cre]
Maintainer:	Kurt Hornik <Kurt.Hornik at R-project.org>
License:	GPL-3
NeedsCompilation:	no
SystemRequirements:	Java (>= 5.0)
In views:	NaturalLanguageProcessing
CRAN checks:	openNLP results

RSentiment: Analyse Sentiment of English Sentences

Analyses sentiment of a sentence in English and assigns score to it. It can classify sentences to the following categories of sentiments:- Positive, Negative, very Positive, very negative, Neutral. sentences, it counts the number of sentences in each category of sentiment.In calculating the score, negation and various degrees of adjectives are taken into consideration. It deals only with En

Version:	2.2.2
Depends:	R (≥ 2.1.0)
Imports:	plyr, stringr, openNLP, NLP
Suggests:	knitr, rmarkdown, testthat

CONCLUSION : LE TDM…

Concerne et impacte :

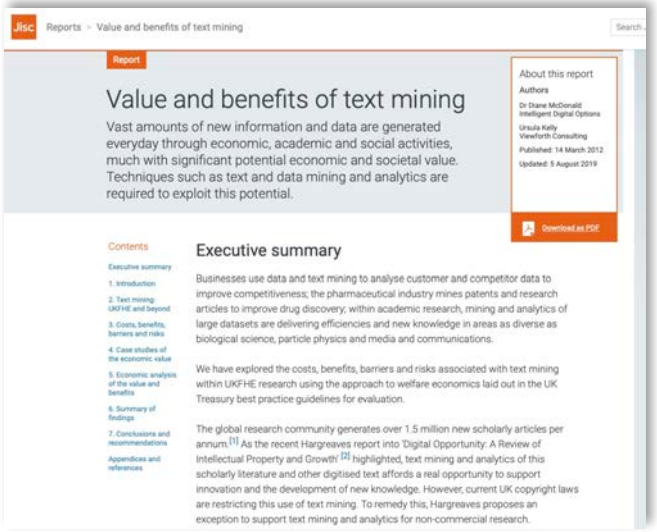
- La recherche scientifique dans son ensemble
- La société au travers d'applications du quotidien

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

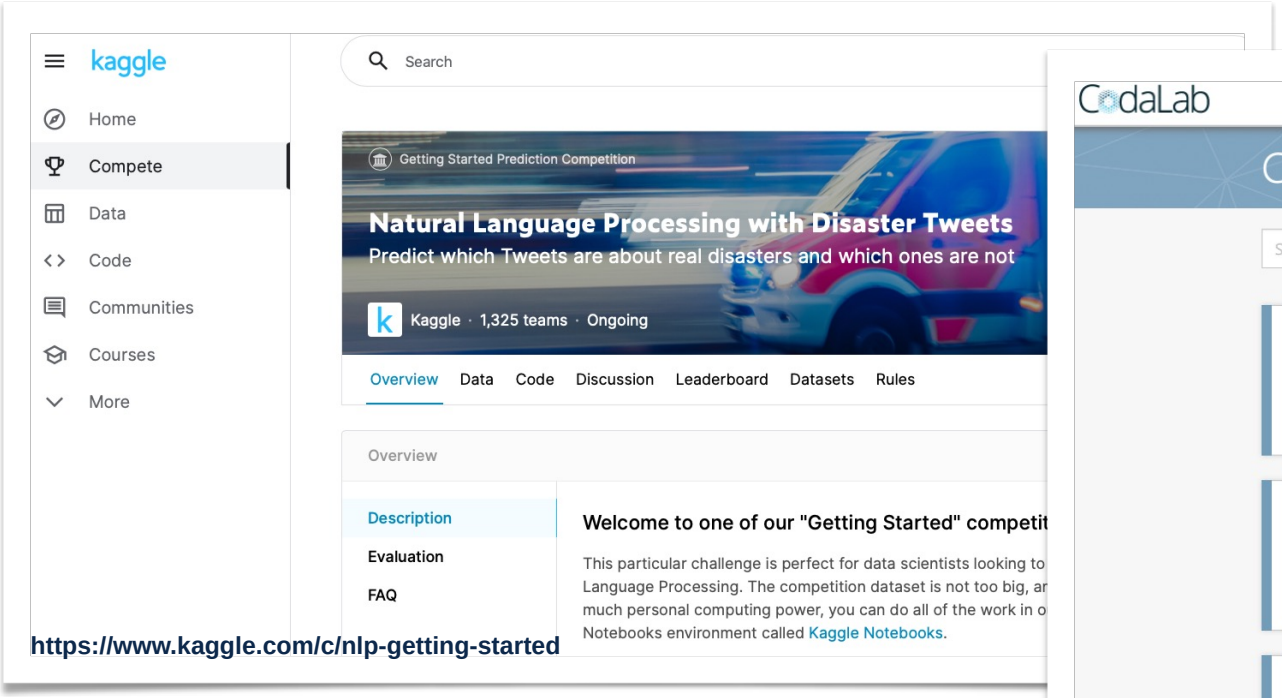
Social factors & opinion

- **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

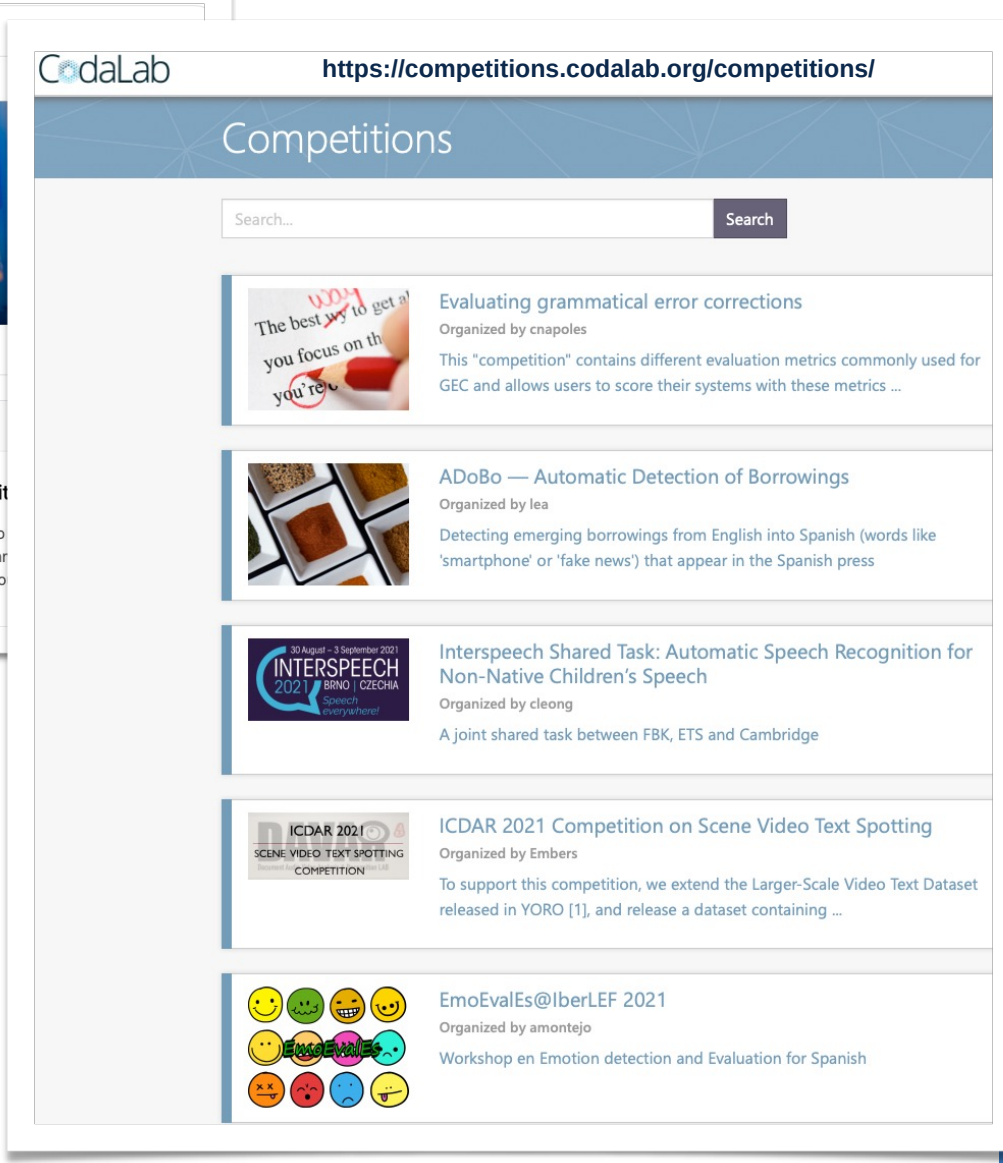
Information in scientific & clinical text

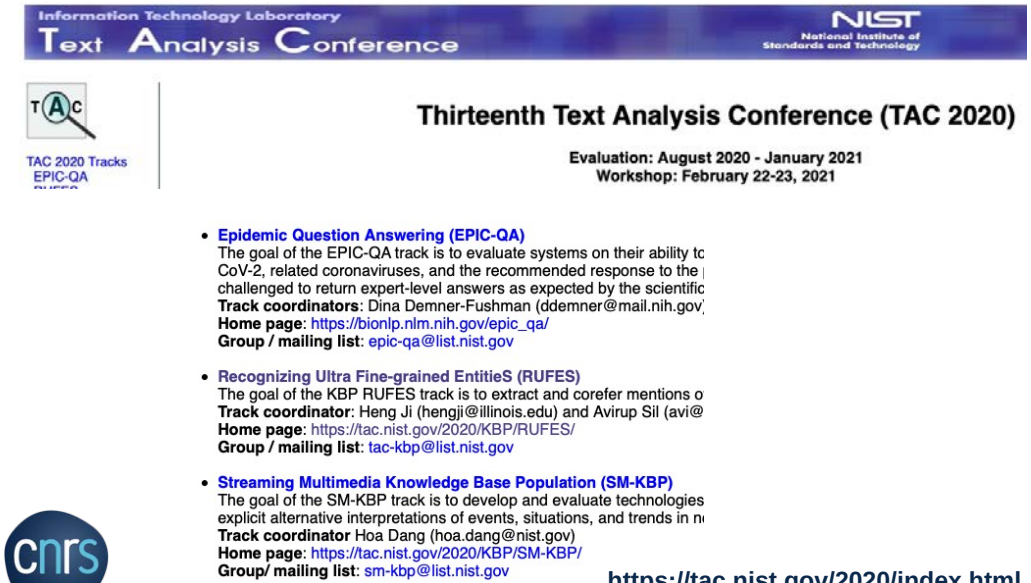
- **Task 8: MeasEval: Counts and Measurements** ([email organizers] [mailing list])
  Corey Harper, Jessica Cox, Ron Daniel, Paul Groth, Curt Kohler, Antony Scerri
- **Task 9: Statement Verification and Evidence Finding with Tables** ([email organizers] [mailing list])
  Nancy Xin Ru Wang, Sara Rosenthal, 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)



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





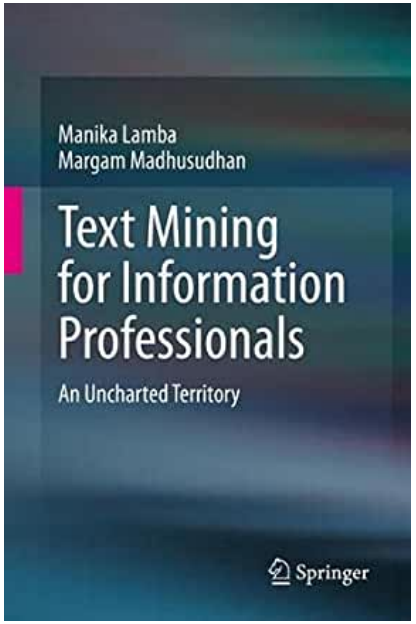
https://tac.nist.gov/2020/index.html

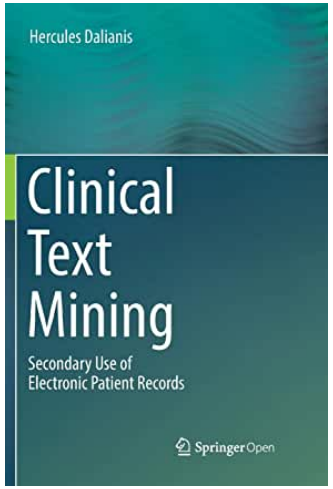
De nombreux ateliers et compétitions (3)

DEFT (DÉfi Fouille de Textes)

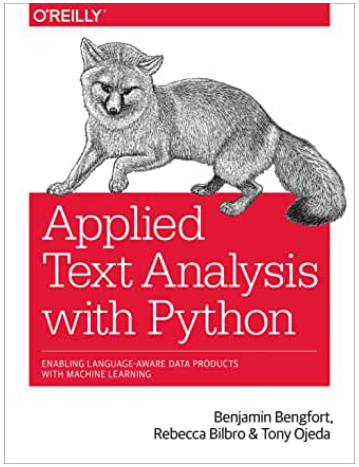
- **2005** *(Dourdan, France, TALN 2005)* : identification du locuteur d'un discours politique parmi deux protagonistes différents (Jacques Chirac vs. François Mitterrand).
- **2006** *(Fribourg, Suisse, SDN 2006)* : segmentation thématique de textes politiques.
- **2007** *(Grenoble, France, AFIA 2007)* : 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)* : recherch
- **2020** *(~~Nancy, France,~~ conférence virtuelle JEP-TALN-RECITAL*
  cliniques

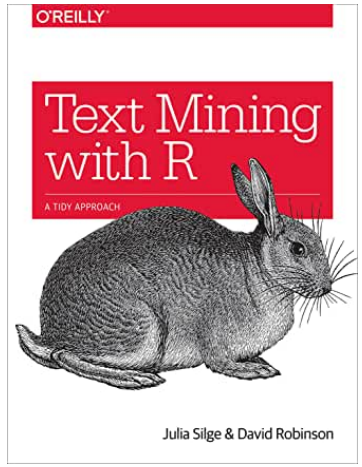
https://deft.lisn.upsaclay.fr/2021/

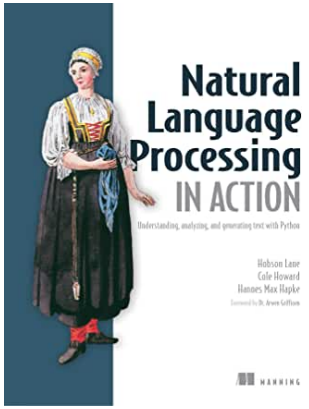

Manika Lamba
Margam Madhusudhan
Text Mining for Information Professionals
An Uncharted Territory
Springer

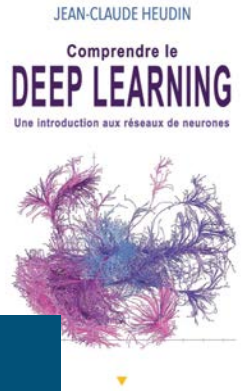

Hercules Dalianis
Clinical Text Mining
Secondary Use of Electronic Patient Records
SpringerOpen

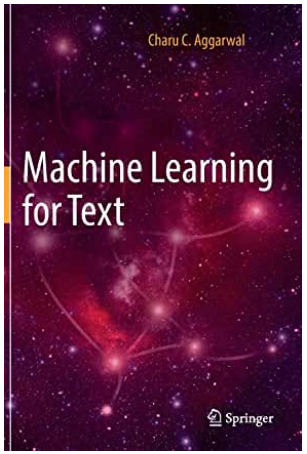

Collection Systèmes d'information et organisations documentaires
sous la direction de Stéphane Chaudiron
Expérimentations et évaluations en fouille de textes
un panorama des campagnes DEFT
sous la direction de
Cyril Grouin
Dominic Forest
Hermes
Lavoisier

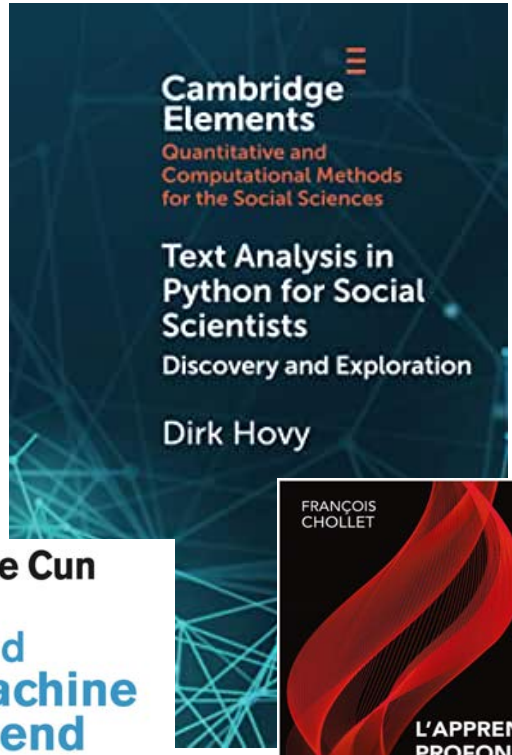

O'REILLY
Applied Text Analysis with Python
ENABLING LANGUAGE-AWARE DATA PRODUCTS WITH MACHINE LEARNING
Benjamin Bengfort,
Rebecca Bilbro & Tony Ojeda

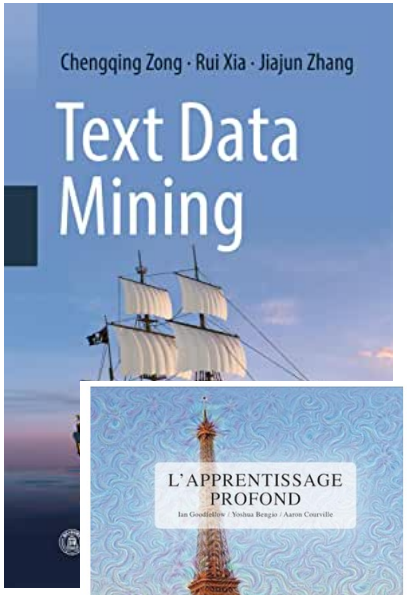

O'REILLY
Text Mining with R
A TIDY APPROACH
Julia Silge & David Robinson


Natural Language Processing IN ACTION
Hobson Lane
Cole Howard
Hannes Max Hapke
MANNING

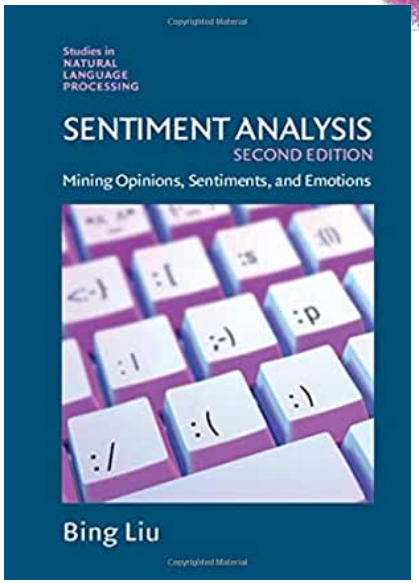

JEAN-CLAUDE HEUDIN
Comprendre le
DEEP LEARNING
Une introduction aux réseaux de neurones

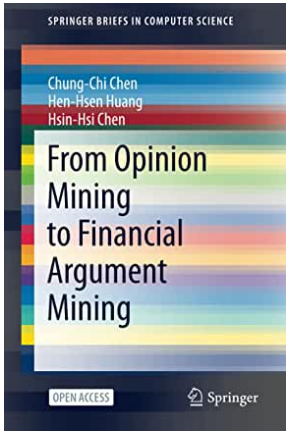

Charu C. Aggarwal
Machine Learning for Text
Springer


Cambridge Elements
Quantitative and Computational Methods for the Social Sciences
Text Analysis in Python for Social Scientists
Discovery and Exploration
Dirk Hovy

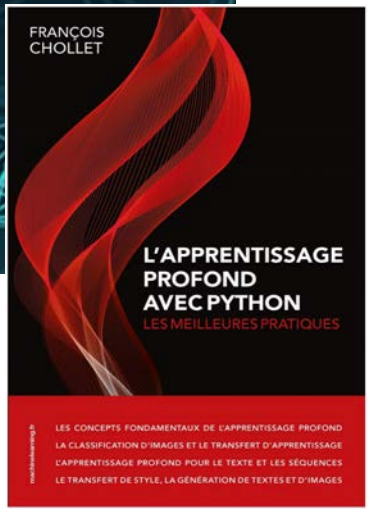

Chengqing Zong · Rui Xia · Jiajun Zhang
Text Data Mining

L'APPRENTISSAGE PROFOND
Ian Goodfellow / Yoshua Bengio / Aaron Courville
MASSOT EDITIONS
Quantmetry


Copyrighted Material
Studies in NATURAL LANGUAGE PROCESSING
SENTIMENT ANALYSIS
SECOND EDITION
Mining Opinions, Sentiments, and Emotions
Bing Liu
Copyrighted Material


SPRINGER BRIEFS IN COMPUTER SCIENCE
Chung-Chi Chen
Hen-Hsen Huang
Hsin-Hsi Chen
From Opinion Mining to Financial Argument Mining
OPEN ACCESS
Springer


Yann Le Cun
Prix Turing
Quand la machine apprend
La révolution des neurones artificiels et de l'apprentissage profond
L'HOMME QUI RÉVOLUTIONNE L'INTELLIGENCE ARTIFICIELLE
Odile Jacob


FRANÇOIS CHOLLET
L'APPRENTISSAGE PROFOND AVEC PYTHON
LES MEILLEURES PRATIQUES
LES CONCEPTS FONDAMENTAUX DE L'APPRENTISSAGE PROFOND
LA CLASSIFICATION D'IMAGES ET LE TRANSFERT D'APPRENTISSAGE
L'APPRENTISSAGE PROFOND POUR LE TEXTE ET LES SÉQUENCES
LE TRANSFERT DE STYLE, LA GÉNÉRATION DE TEXTES ET D'IMAGES