What is NLP?

Yogesh Kulkarni

July 25, 2022

Use case: Bank Call Center

Calling the Call Center

- Calling to an IVR (Integrated voice response)
- A prerecorded menu selection.
- "Please press 1 for Account Details, Please press 2 for . . . "
- till it comes to your option.
- towards end, somewhere, given access to a person to talk to.

Boring? Annoying?

(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

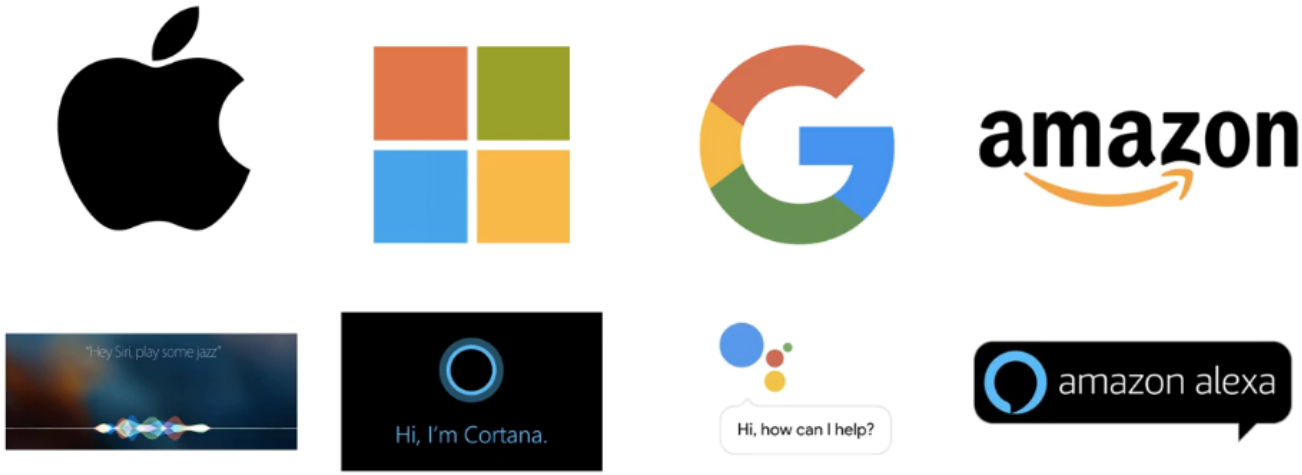
Instead, how about typing/saying your query directly and getting the answer right away?

Solution

Chatbots

- Which problem of IVR it is solving?
- Advantages?
- Disadvantages?
- Gaining popularity . . .
- Many platforms
- Companies in Pune?

The Giants are at it ...



(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

- Chatbots or QA systems, predominantly voice based,
- Underlying processing is primarily Natural Language Processing (NLP).
- You can have your own chatbot, specific to you!!
- NLP is the core skill needed.

Why so much popularity?

Chatbots are:

- Autonomous and Always Available
- Drive Conversation
- Able to handle millions of requests, scalable.

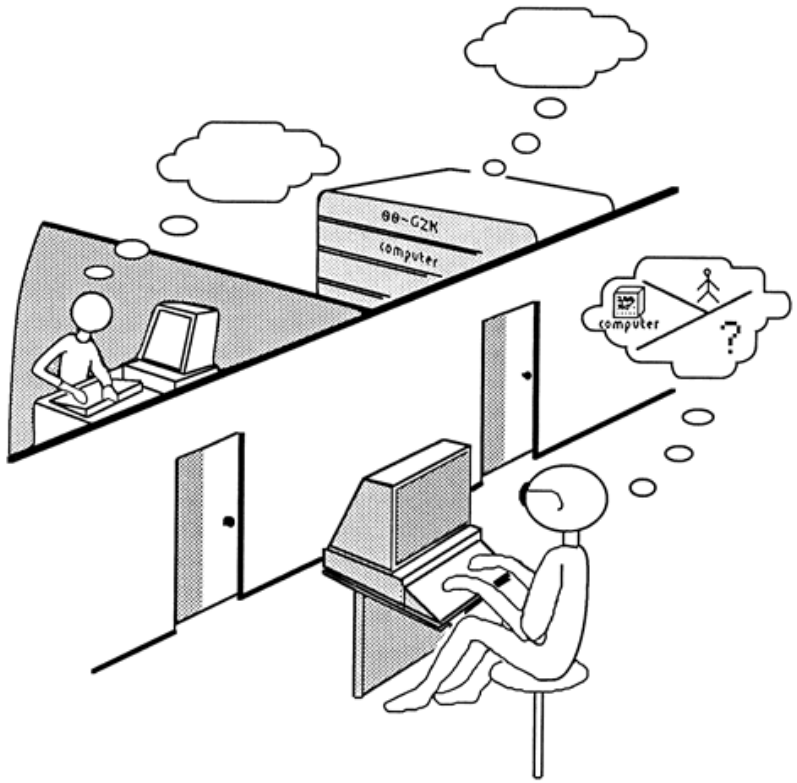
Its hard to master language, and thus NLP.

NLP is AI-complete

- "The most difficult problems in AI manifest themselves in human language phenomena."
- Use of language is the touchstone of intelligent behavior.
- Test for Intelligence - Turing Test
- Alan Turing (1950) proposed a test of a machine's capability to perform human-like conversation.

Turing Test

A human judge engages in a natural language conversation with two other parties, one a human and the other a machine; if the judge cannot reliably tell which is which, then the machine is said to pass the test.



Early Conversational Programs

- ELIZA (by Joseph Weizenbaum), 1966.
- A psychotherapist, but NO real understanding;
- Simple pattern-matching to respond to user input to canned responses

(my ?x depresses me) (why does your ?x depress you)
(life ?x) (why do you say it ?x)
(I could ?x) (you could ?x)
(because ?x) (that is a good reason)
(?x) (tell me more)

```
>> Hello.
How do you do.  Please state your problem.
>> Men are all alike.
In what way ?
>> They're alwas bugging us about something or other.
I'm not sure I understand you fully.
>> They're always bugging us about something or other.
Can you think of a specific example ?
>> Well, my boyfriend made me come here.
Your boyfriend made you come here  ?
>> He says I'm depressed much of the time.
I am sorry to hear that you are depressed.
```

Loebner Prize

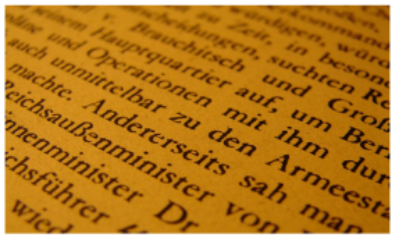
- In 1990, Hugh Loebner started Turing Test competition
- $100,000 will be awarded to the first bot that judges cannot distinguish from a real human in a Turing test that includes text, visual, and auditory input.
- Nobody has won the grand prize yet.
- 2016 (and 2013) year-wise top winner - Mitsuku. https://www.facebook.com/mitsukubot

Why can't we win the Grand Prize? What are the challenges? Why Language is hard? What is Language?

What is Language?

Language Types

Natural Language



(http://expertenough.com/2392/german-language-hacks)

日本語で

冬は世界各地でさまざまなお祝いが行われる時期です。ほんのいくつか例を挙げるだけでも、ハナカ、クリスマス、クワンザ、新年などさまざまなお祝いがあります。各文化によってその祝い方はさまざまですが、ほとんどのお祝いにはごちそうが欠かせません。

(http://www.transparent.com/learn-japanese/articles/dec_99.html)

Artificial Language

```
try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.g
            acWriter.println(cMessage);
            acWriter.flush();
        } catch(IO
            System
            queue.
        }
    }
} catch(Interrupte
```

append(CharSequenc
append(char c)
append(CharSequenc
format(String form
format(Locale l, S
printf(String form

(https://netbeans.org/features/java/)

```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '    %s [label="%s" % (nodename,label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s"];' % ast[1]
        else:
            print '"]'
    else:
        print '"];'
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print '    %s -> {' % nodename,
        for name in children:
            print '%s' % name,
```

(http://noobite.com/learn-programming-start-with-python/)

Differences?

Language, simplistically

- A vocabulary consists of a set of words
- A text is composed of a sequence of words from a vocabulary
- A language is constructed of a set of all possible texts

(http://learnenglish.britishcouncil.org/en/vocabulary-games)

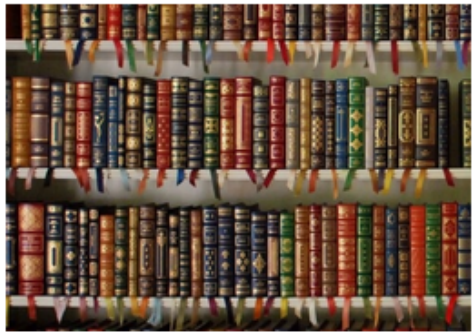
THIS WEEK

EDITORIALS

Beyond the genome

Studies of the epigenetic signatures of many healthy and diseased human tissues could provide crucial information to link genetic variation and disease.

(http://www.nature.com/polopoly_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf)



(http://www.old-engli.sh/language.php)

NLP

- NLP is Natural Language Processing, ie processing Natural Langauge for some end-purpose in mind.
- Inspite of usage of Natural Language for thousands of years, why are we not able to process it well?

NLP Challenges

Paraphrasing

Paraphrasing: Different words/sentences express the same meaning

- Season of the year: Fall/Autumn
- Book delivery time
  - When will my book arrive?
  - When will I receive my book?

Ambiguity

Ambiguity: One word/sentence can have different meanings

- Fall
  - The third season of the year
  - Moving down towards the ground or towards a lower position
- The door is open
  - Expressing a fact
  - A request to close the door

Syntax and ambiguity

"I saw the man with a telescope."
- Who had the telescope?

Semantics

The astronomer loves the star.

- Star in the sky
- Celebrity

(http://en.wikipedia.org/wiki/Star#/media/File:Starsinthesky.jpg)

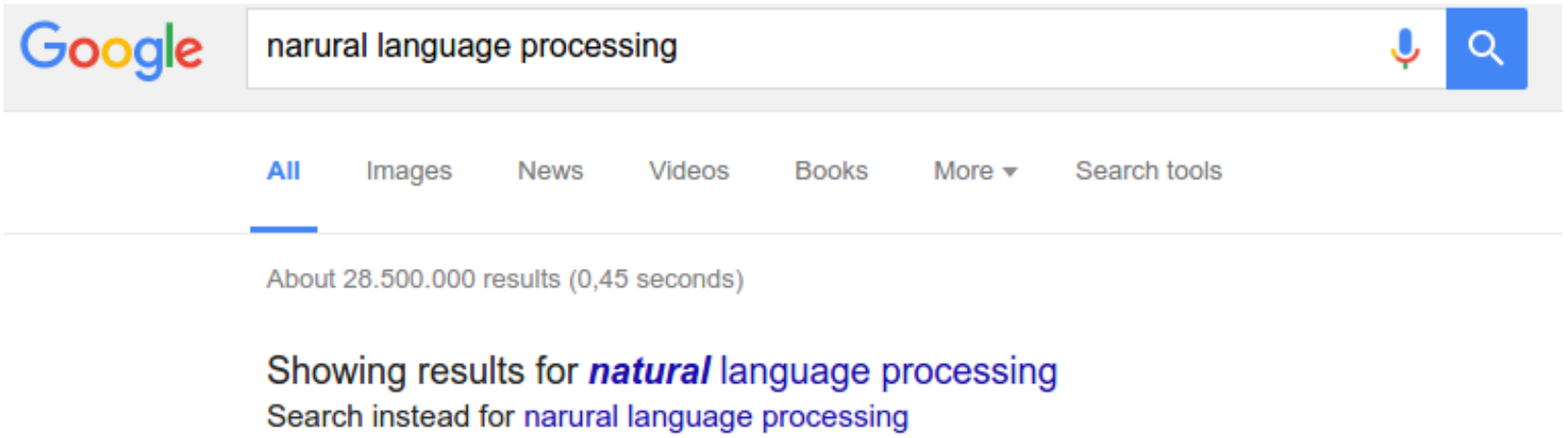
(http://www.businessnewsdaily.com/2023-celebrity-hiring.html)

NLP Applications

Grammar

Spell and Grammar Checking

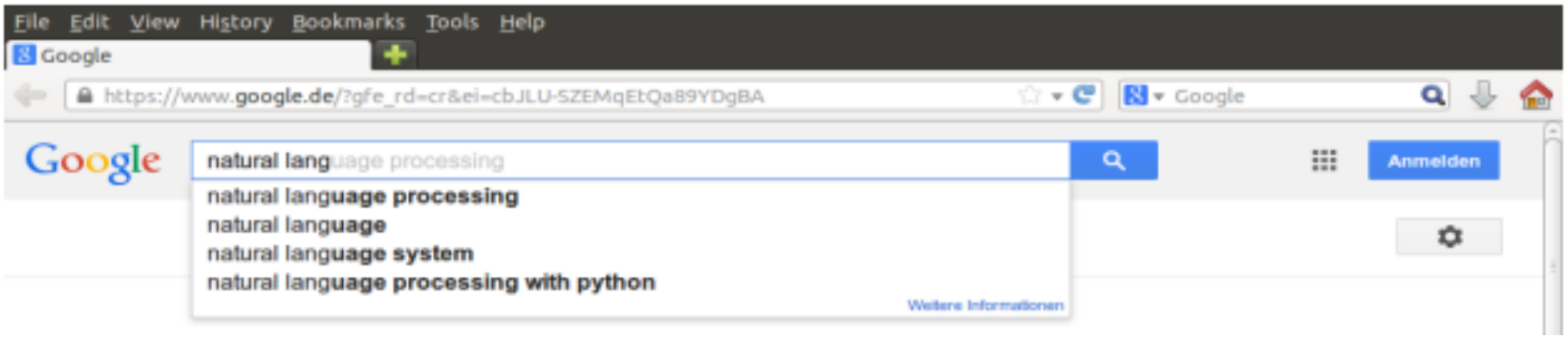
- Checking spelling and grammar
- Suggesting alternatives for the errors



Word Prediction

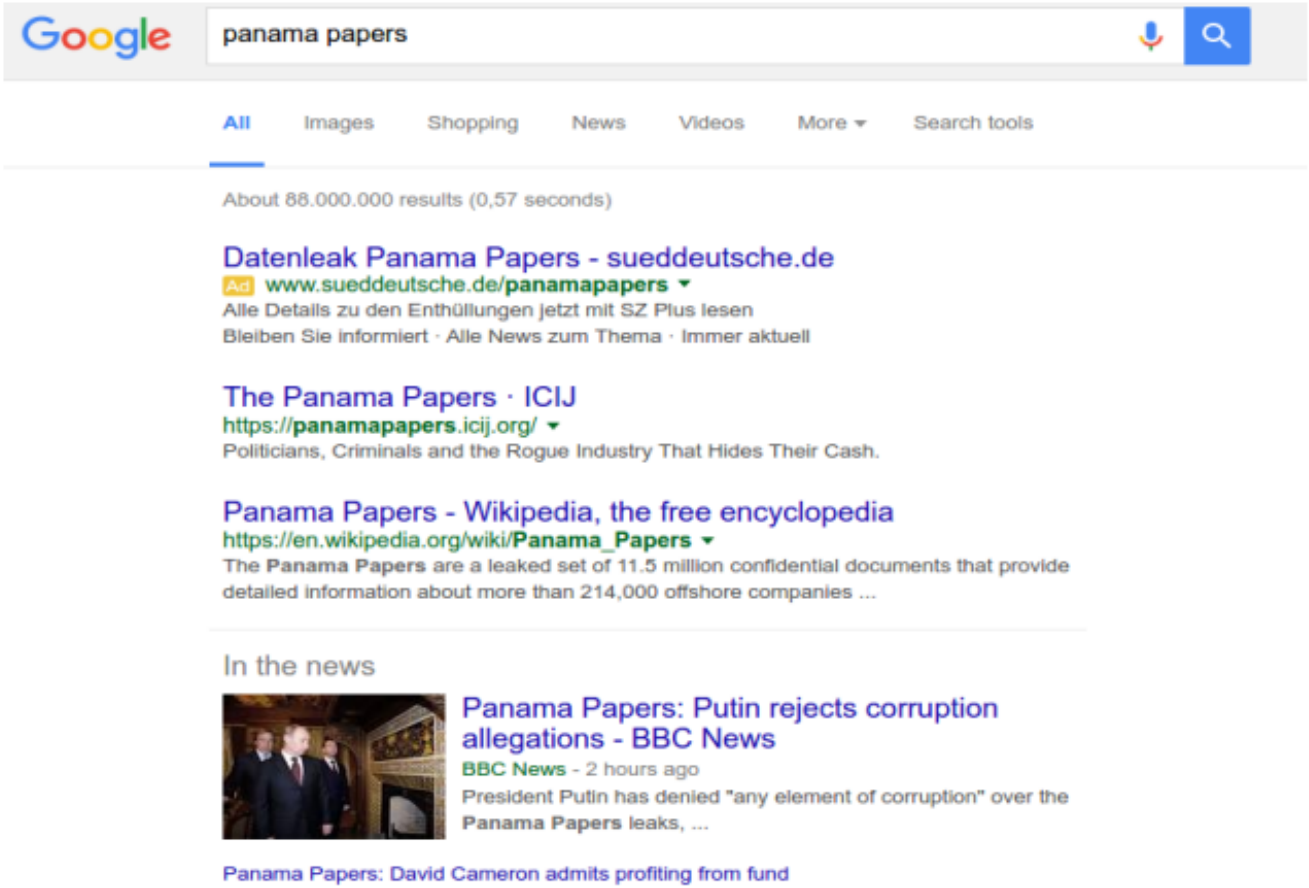
Word Prediction: Predicting the next word that is highly probable to be typed by the user

- Mobile typing
- Search Engines



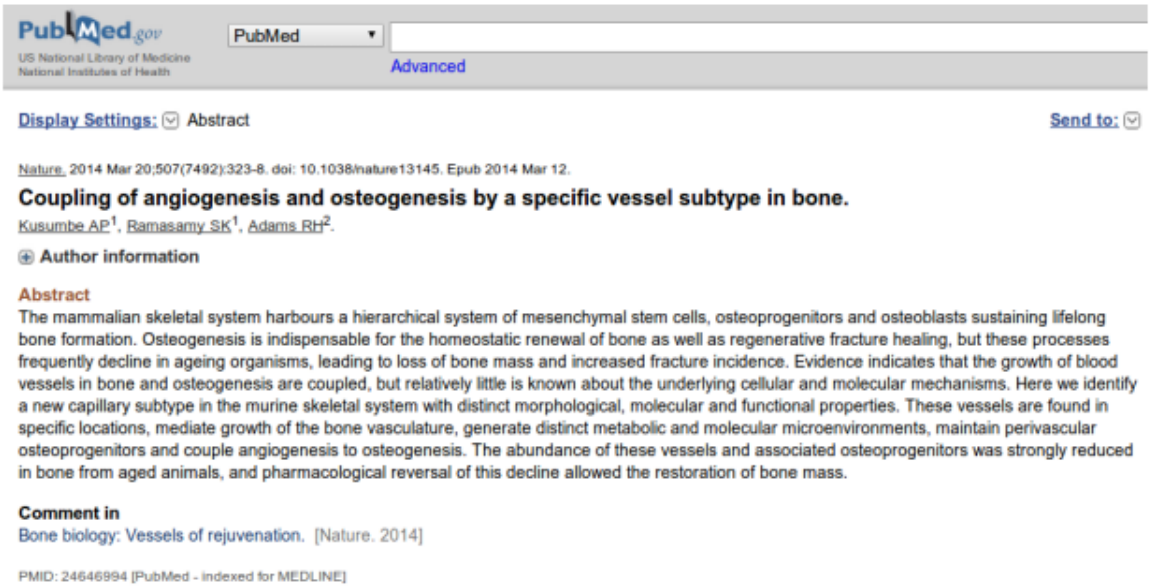
Information Retrieval

Information Retrieval: Finding relevant information to the user's query



Text Categorization

Text Categorization: Assigning one (or more) pre-defined category to a text



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Nature. 2014 Mar 20;507(7492):323-8. doi: 10.1038/nature13145. Epub 2014 Mar 12.

Coupling of angiogenesis and osteogenesis by a specific vessel subtype in bone.

Kusumbe AP[1], Ramasamy SK[1], Adams RH[2].

Author information

Abstract

The mammalian skeletal system harbours a hierarchical system of mesenchymal stem cells, osteoprogenitors and osteoblasts sustaining lifelong bone formation. Osteogenesis is indispensable for the homeostatic renewal of bone as well as regenerative fracture healing, but these processes frequently decline in ageing organisms, leading to loss of bone mass and increased fracture incidence. Evidence indicates that the growth of blood vessels in bone and osteogenesis are coupled, but relatively little is known about the underlying cellular and molecular mechanisms. Here we identify a new capillary subtype in the murine skeletal system with distinct morphological, molecular and functional properties. These vessels are found in specific locations, mediate growth of the bone vasculature, generate distinct metabolic and molecular microenvironments, maintain perivascular osteoprogenitors and couple angiogenesis to osteogenesis. The abundance of these vessels and associated osteoprogenitors was strongly reduced in bone from aged animals, and pharmacological reversal of this decline allowed the restoration of bone mass.

Comment in

Bone biology: Vessels of rejuvenation. [Nature. 2014]

PMID: 24646994 [PubMed - indexed for MEDLINE]

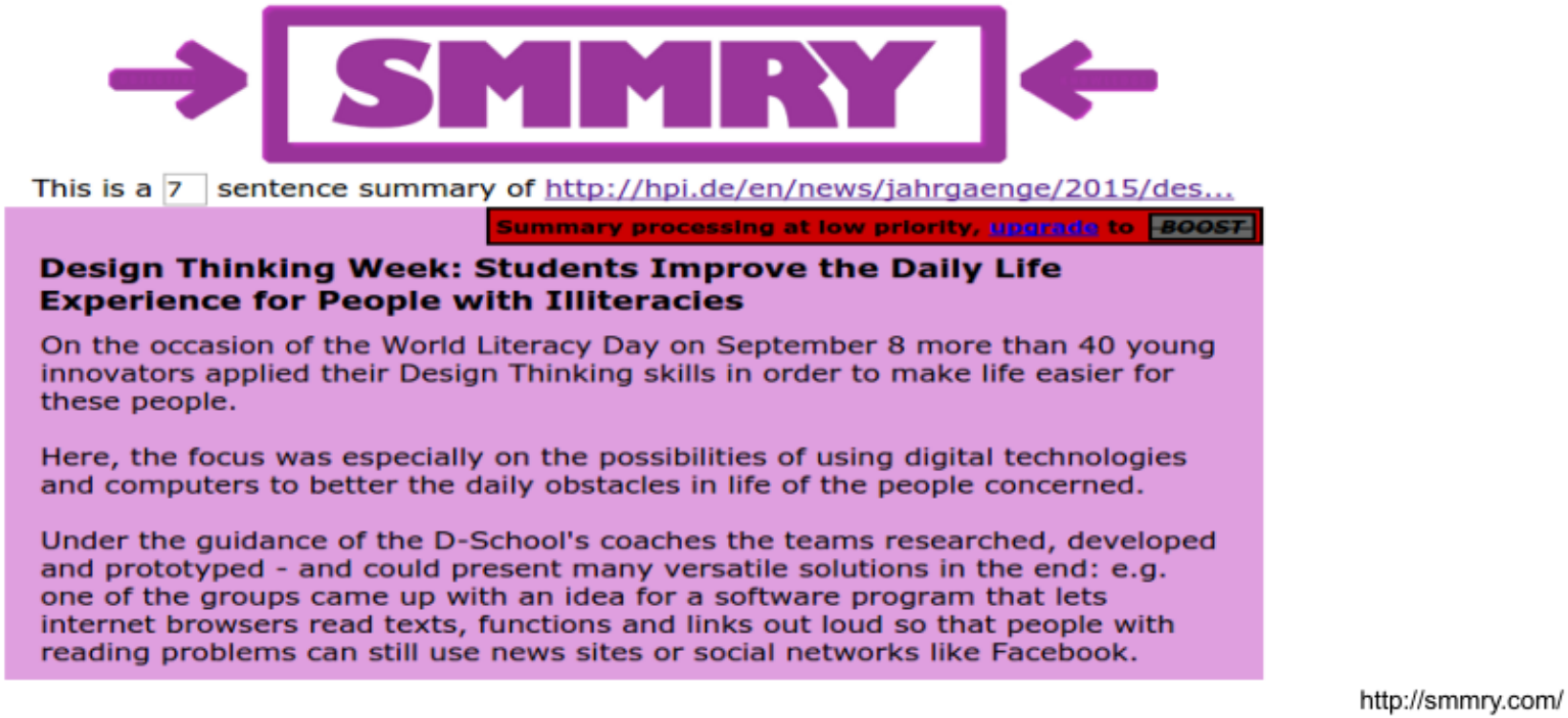
MeSH Terms

Aging/metabolism
Aging/pathology
Animals
Blood Vessels/anatomy & histology
Blood Vessels/cytology
Blood Vessels/growth & development
Blood Vessels/physiology*
Bone and Bones/blood supply*
Bone and Bones/cytology
Endothelial Cells/metabolism
Hypoxia-Inducible Factor 1, alpha Subunit/metabolism
Male
Mice
Mice, Inbred C57BL
Neovascularization, Physiologic/physiology*
Osteoblasts/cytology
Osteoblasts/metabolism
Osteogenesis/physiology*
Oxygen/metabolism
Stem Cells/cytology
Stem Cells/metabolism

Text Categorization

Summarization

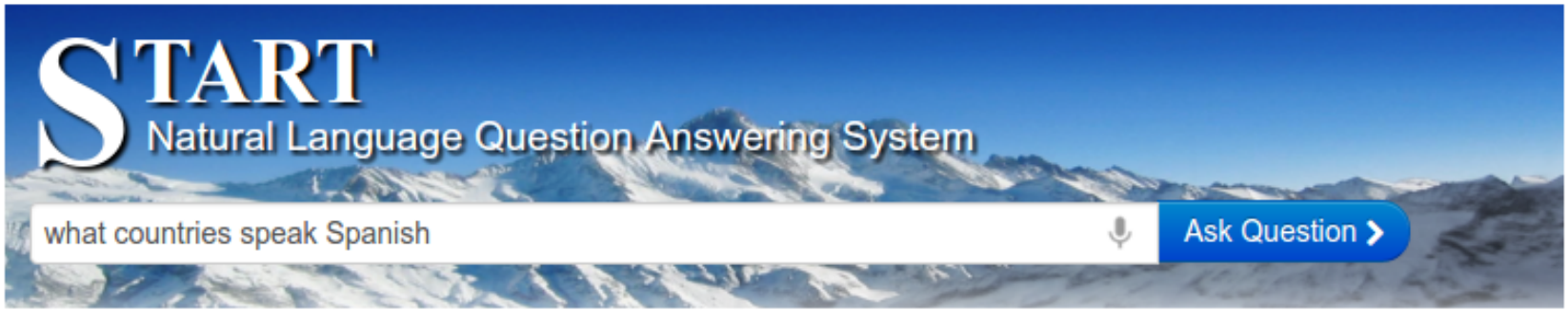
Summarization: Generating a short summary from one or more documents, sometimes based on a given query



http://smmry.com/

Question answering

Question answering: Answering questions with a short answer



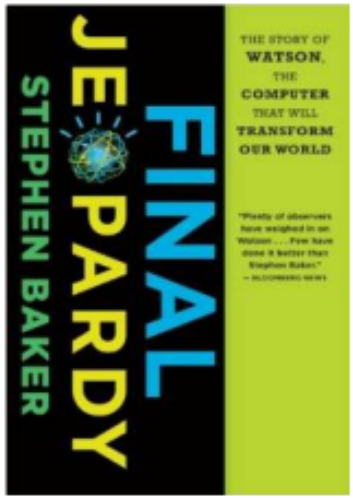
===> what countries speak Spanish

The language Spanish is spoken in Argentina, Aruba, Belize, Bolivia, Brazil, Canada, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Falkland Islands (Islas Malvinas), Gibraltar, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Martin, Sint Maarten, Spain, Switzerland, Trinidad and Tobago, United States, Uruguay, Venezuela, and Virgin Islands.

The language Castilian Spanish is spoken in Spain.

Question answering

Question answering: IBM Watson in Jeopardy



Information Extraction

Information Extraction: Extracting important concepts from texts and assigning them to slot in a certain template

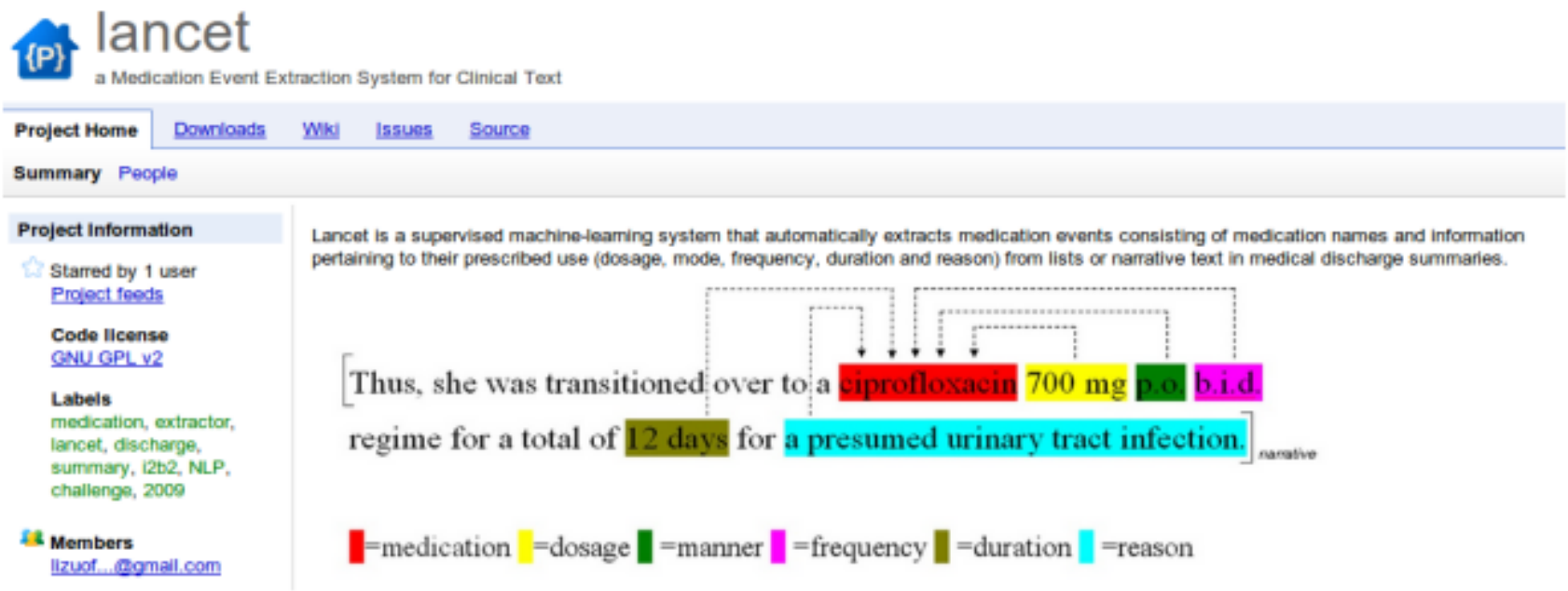
Merkel at the EPP Summit, March 2016

Chancellor of Germany	
Incumbent	
Assumed office	22 November 2005
President	Horst Köhler Christian Wulff Joachim Gauck
Deputy	Franz Müntefering Frank-Walter Steinmeier Guido Westerwelle Philipp Rösler Sigmar Gabriel
Preceded by	Gerhard Schröder
Leader of the Christian Democratic Union	
Incumbent	
Assumed office	

In office	17 November 1994 – 26 October 1998
Chancellor	Helmut Kohl
Preceded by	Klaus Töpfer
Succeeded by	Jürgen Trittin
Minister for Women and Youth	
In office	18 January 1991 – 17 November 1994
Chancellor	Helmut Kohl
Preceded by	Ursula Lehr
Succeeded by	Claudia Nolte
Personal details	
Born	Angela Dorothea Kasner 17 July 1954 (age 61) Hamburg, West Germany
Political party	Democratic Awakening (1989–1990) Christian Democratic Union (1990–present)
Spouse(s)	Ulrich Merkel (1977–1982) Joachim Sauer (1998–present)
Alma mater	Leipzig University
Religion	Lutheranism (within Evangelical Church)
Signature	

Information Extraction

Information Extraction: Includes named-entity recognition



Machine Translation

Machine Translation: Translating a text from one language to another

Sentiment Analysis

Sentiment Analysis: Identifying sentiments and opinions stated in a text

Customer Reviews

Speech and Language Processing, 2nd Edition

15 Reviews

5 star: (8)
4 star: (3)
3 star: (3)
2 star: (0)
1 star: (1)

Average Customer Review

(15 customer reviews)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

4 of 4 people found the following review helpful

Great introductions and reference book

I read the first edition of that book and it is terrific. The second edition is much more adapted to current research. Statistical methods in NLP are more detailed and some syntax-based approaches are presented. My specific interest is in machine translation and dialogue systems. Both chapters are extensively rewritten and much more elaborated. I believe this book is...

Read the full review ›

Published on August 9, 2008 by carheg

› See more 5 star, 4 star reviews

Vs.

The most helpful critical review

37 of 37 people found the following review helpful

Good description of the problems in the field, but look elsewhere for practical solutions

The authors have the challenge of covering a vast area, and they do a good job of highlighting the hard problems within individual sub-fields, such as machine translation. The availability of an accompanying Web site is a strong plus, as is the extensive bibliography, which also includes links to freely available software and resources.

Now for the...

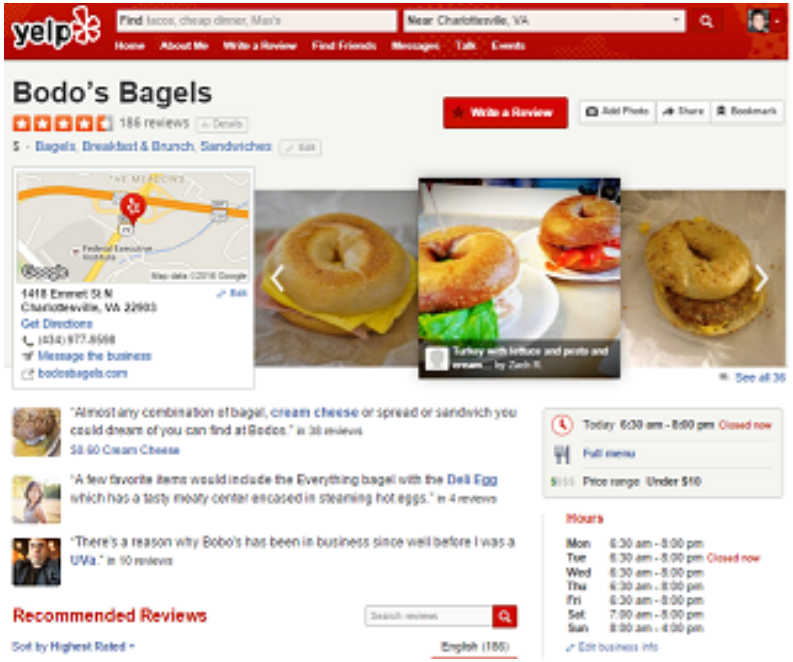
Read the full review ›

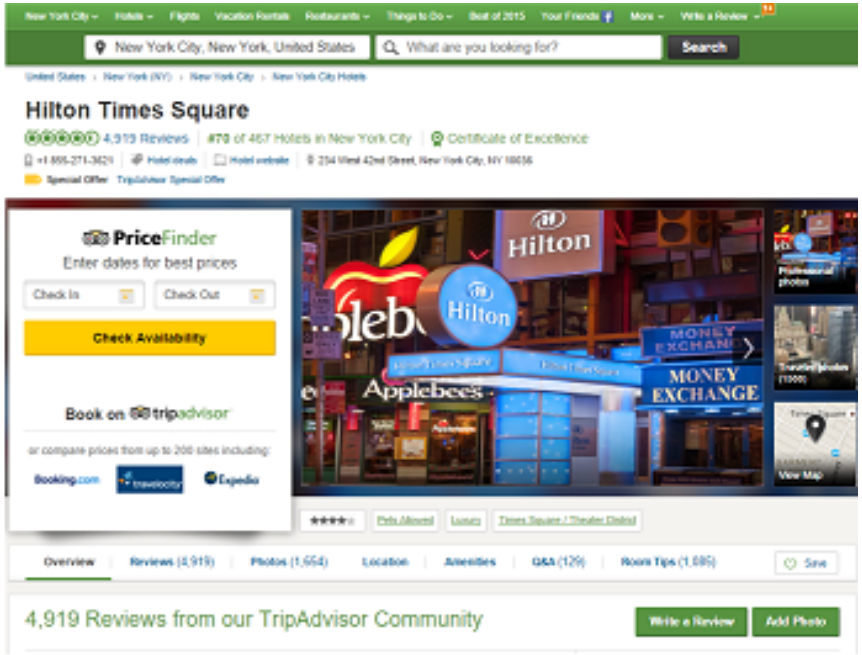
Published on April 2, 2009 by P. Nadkarni

› See more 3 star, 2 star, 1 star reviews

Sentiment Analysis

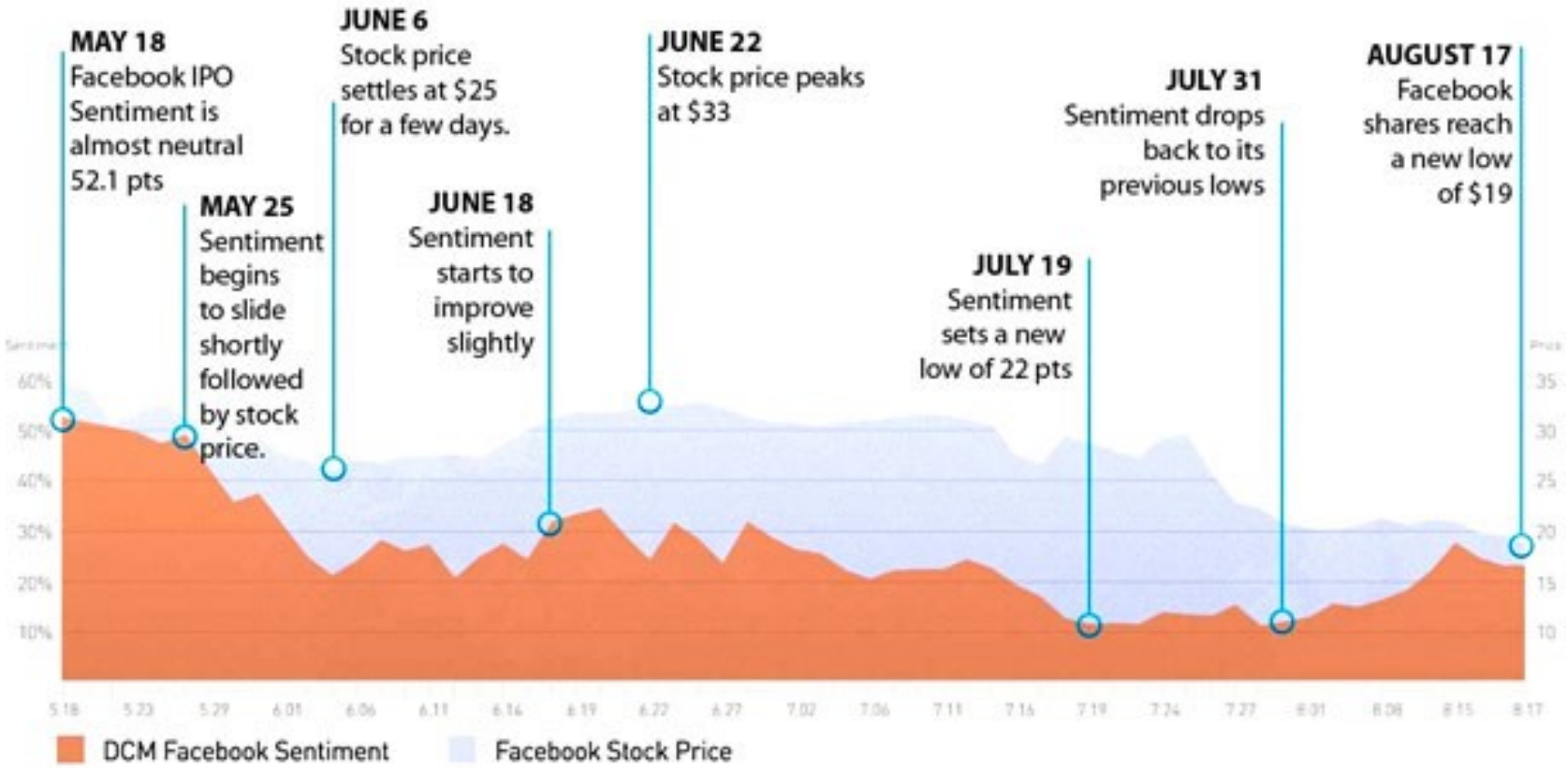
Restaurant/hotel recommendation, Product reviews





Sentiment Analysis

Text analytics in financial services



NLP in today's time

Trends:

- An enormous amount of information is now available in machine readable form as natural language text (newspapers, web pages, medical records, financial filings, product reviews, discussion forums, etc.)
- Conversational agents are becoming an important form of human-computer communication
- Much of human-human interaction is now mediated by computers via social media

Collectively, this means that copious data is available to be used in the development of NLP systems.

Level of difficulties

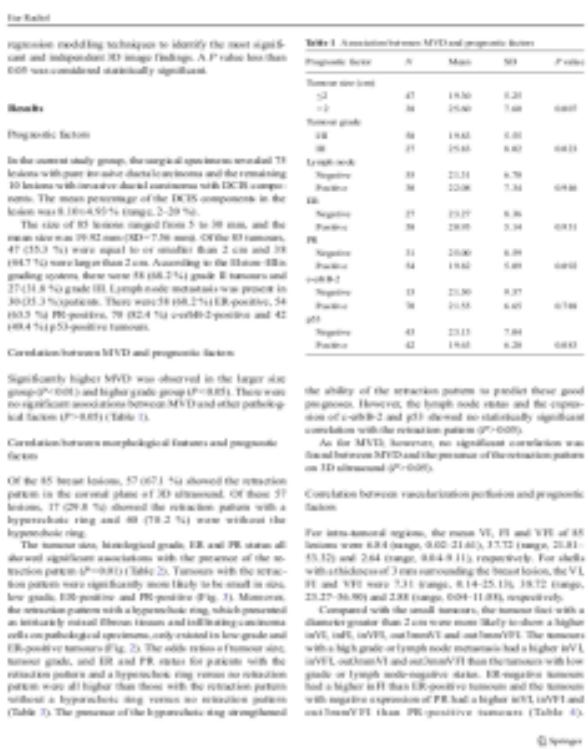
- Easy (mostly solved)
  - Spell and grammar checking
  - Some text categorization tasks
  - Some named-entity recognition tasks
- Intermediate (good progress)
  - Information retrieval
  - Sentiment analysis
  - Machine translation
  - Information extraction
- Difficult (still hard)
  - Question answering
  - Summarization
  - Dialog systems

NLP Activities: How to process text?

Document/Section splitting

Document/Section splitting: Splitting a text into sections

Correlation between three-dimensional ultrasound features and pathological prognostic factors in breast cancer



Sentence splitting

Sentence splitting: Splitting a text into sentences

11 Sentences (= "T-" or "Terminable" units *only* if independent clauses are puctuated as separate sentences, e.g. "I came and he went"-->"I came. And he went.")
Average 23.55 words (SD=12.10)

OBJECTIVES: To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinoma.

METHODS: Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Morphology features and vascularization perfusion on 3D ultrasound were evaluated.

Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c erbB-2 and p53 expression, and microvessel density (MVD) were determined.

Correlations of 3D ultrasound features and prognostic factors were analysed.

RESULTS: The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size (P #8201;= 0.014), a lower histological grade (P #8201;= 0.009) and positive ER or PR expression status (P #8201;= 0.001, 0.044).

The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours.

The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer.

The increased intra-tumour vascularization index (VI, the mean tumour vascularity) reflected a higher histological grade (P #8201;= 0.025) and had a positive correlation with MVD (r #8201;= 0.530, P #8201;= 0.001).

CONCLUSIONS: The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

KEY POINTS: • Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer. • The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis. • The increased intra-tumour vascularization index reflects a higher histological grade. • The intra-tumour vascularization index is positively correlated with microvessel density.

Built by Lextutor, SD function added 18 Nov 2010

Tokenization

Tokenization

- Process of breaking a stream of text up into tokens (= words, phrases, symbols, or other meaningful elements)
- Typically performed at the "word" level
- Not easy: Hewlett-Packard, U.S.A., in some languages there is no "space" between words!

Stemming

Stemming

- Reduces similar words to a given "stem"
- E.g. detects, detected, detecting, detect : detect (stem).
- Usually set of rules for suffix stripping
- Most popular for English: Porter's Algorithm
- 36% reduction in indexing vocabulary (English)
- Linguistic correctness of resulting stems not necessary (sensitivities : sensit)

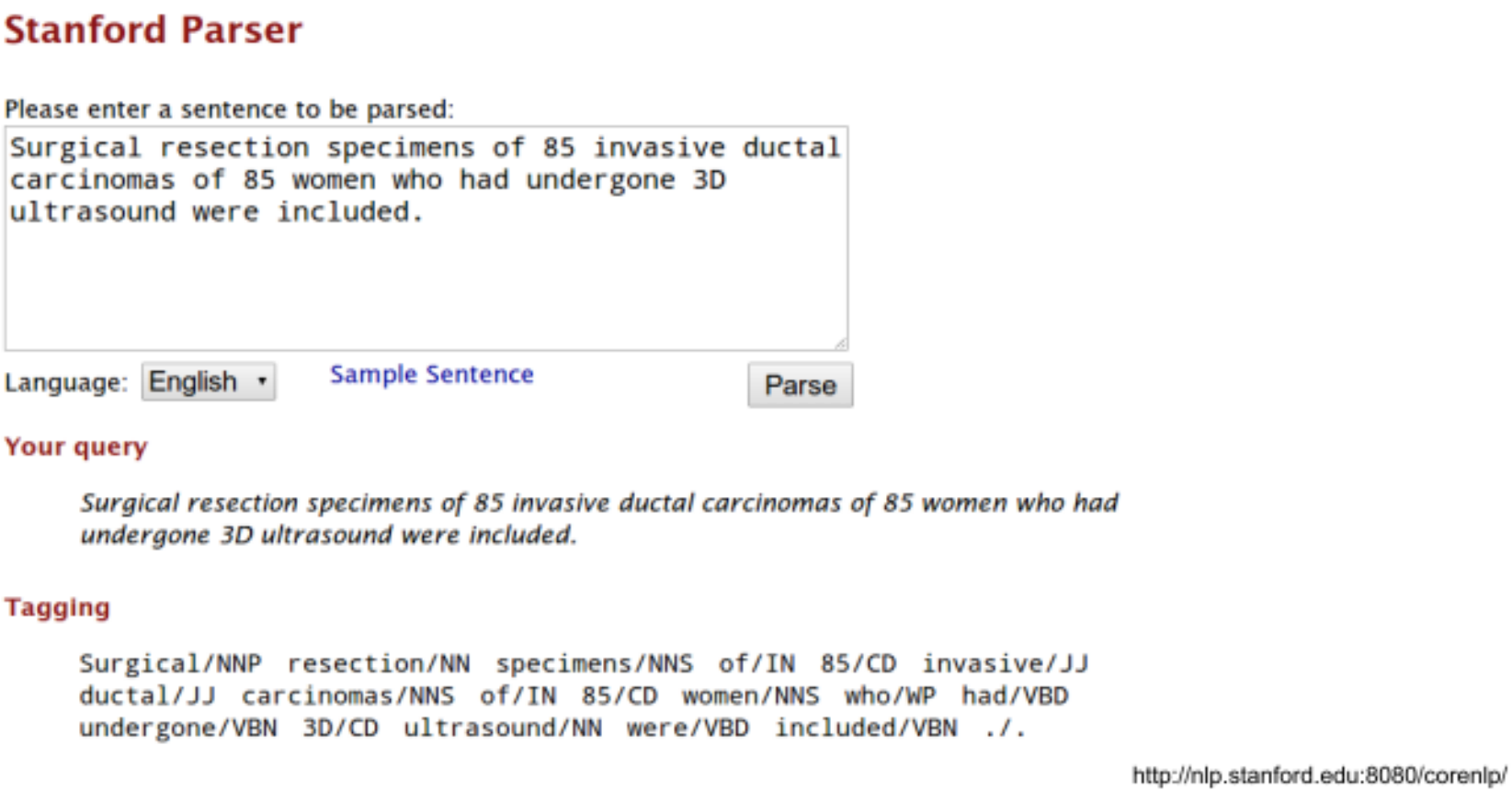
Lemmatization

Lemmatization

- Uses a vocabulary and full morphological analysis of words
- Aims to remove inflectional endings only
- Return the base or dictionary form of a word, which is known as the lemma.
- E.g. saw : see,
  been, was : be

Part-of-speech tagging

Part-of-speech tagging: Assigning a syntatic tag to each word in a sentence



Parsing

Parsing: Building the syntactic tree of a sentence

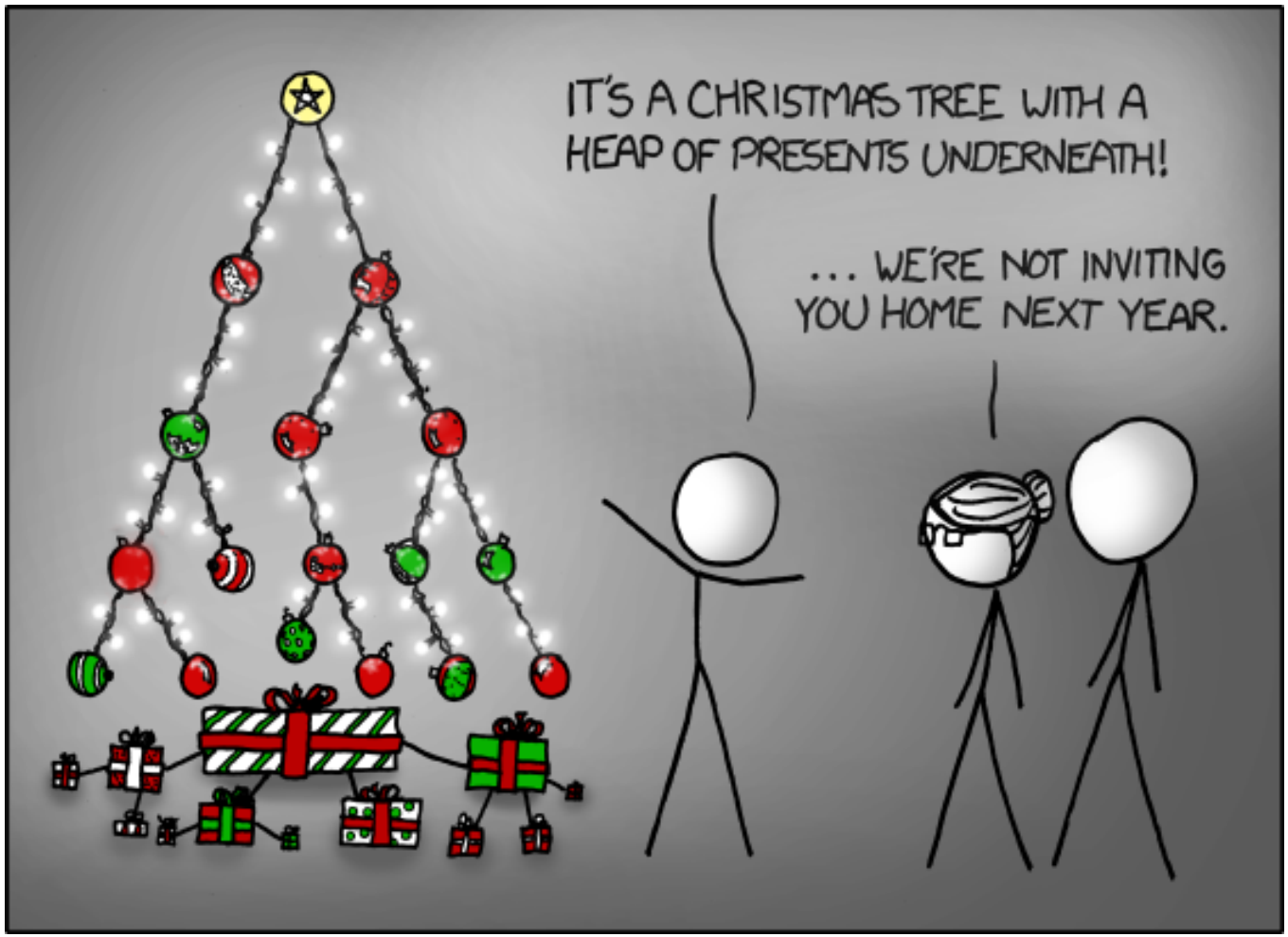
Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound))))))))))))
    (VP (VBD were)
      (VP (VBN included)))
    (. .)))
```

http://nlp.stanford.edu:8080/corenlp/

Parsing

$((DaimlerChryslersshares)_{NP}(rose(threeeights)_{NUMP}(to22)_{PP-NUM})_{VP})_{S}$



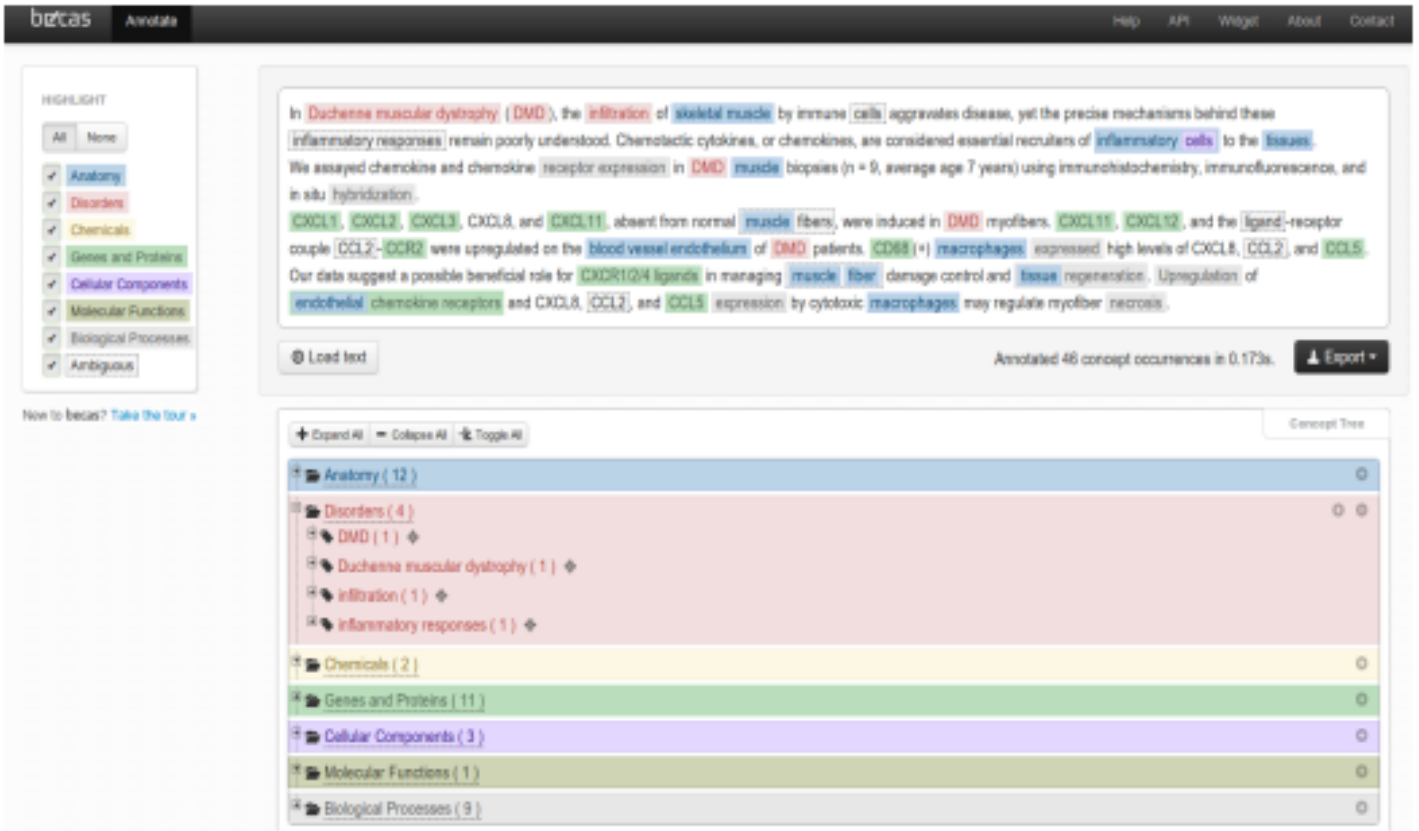
Syntax Tree

Syntax: Sample English grammar

$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid TWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

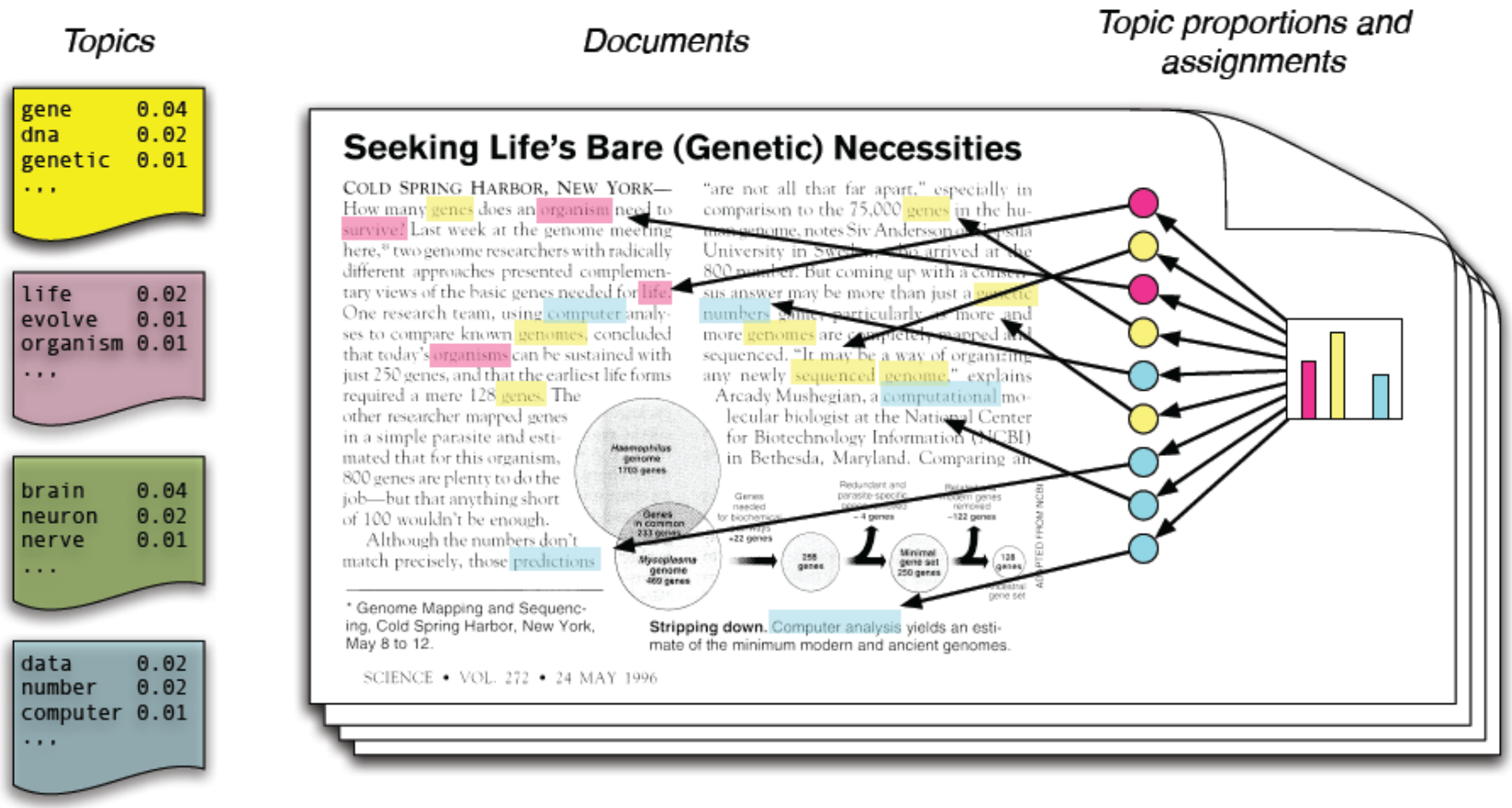
Named-entity recognition

Named-entity recognition: Identifying pre-defined entity types in a sentence



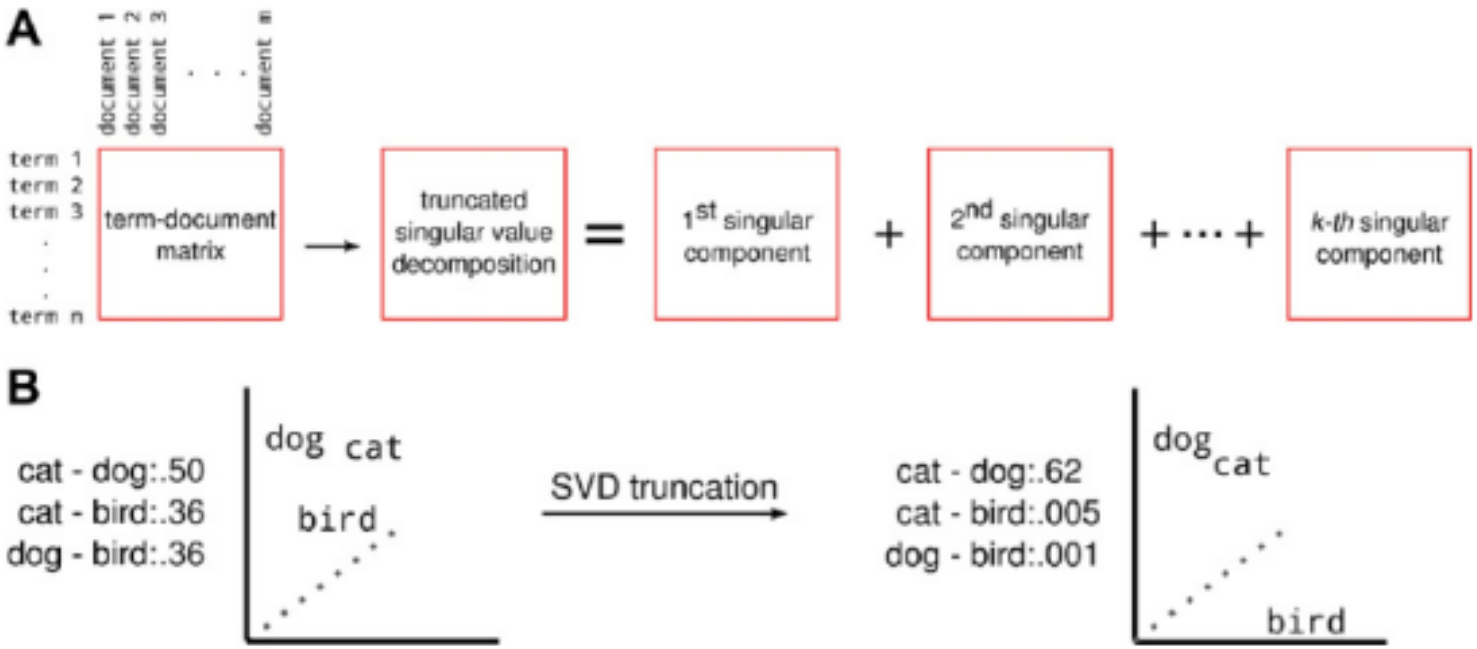
Topic modelings

Topic modeling: Identifying structures in the text corpus



Word embeddings

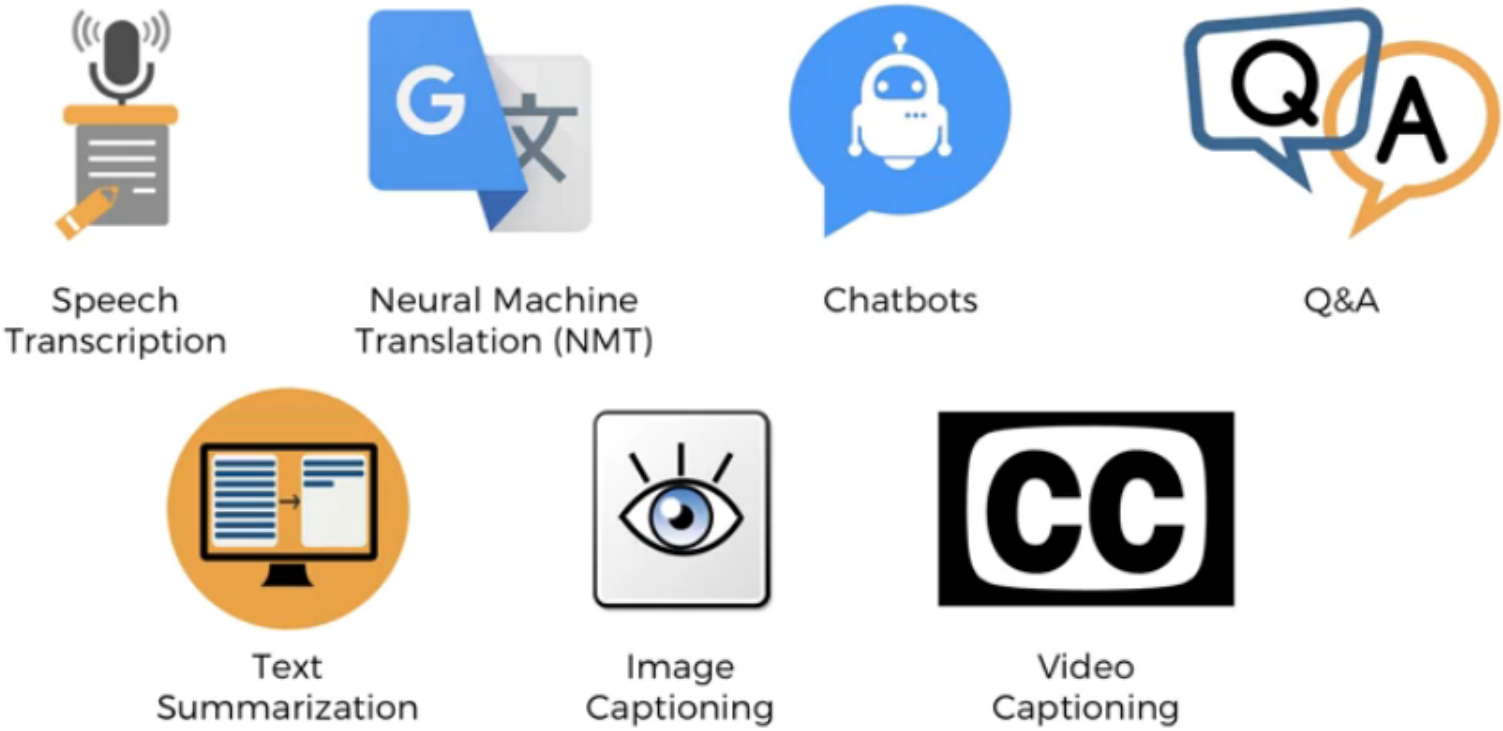
Word embeddings: Compute a vector representing the distributed representation for every word



What next?

- Coursera : Dr Radev's NLP course (https://www.coursera.org/learn/natural-language-processing)
- Course: Deep NLP By Richard Sochar (Stanford)
- Book: Natural Language Processing with Python

NLP Opportunities



(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

References

Many publicly available resources have been refereed for making this presentation. Some of the notable ones are:

- Introduction to Natural Language Processing - Dr. Mariana Neves, SoSe 2016
- Machine Learning for Natural Language Processing - Traian Rebedea, Stefan Ruseti - LeMAS 2016 - Summer School
- CSC 594 Topics in AI - Natural Language Processing - De Paul
- Deep Learning for Natural Language Processing - Sihem Romdhani
- Notebooks and Material @ https://github.com/rouseguy/DeepLearningNLP_Py

Thanks ... yogeshkulkarni@yahoo.com