What is NLP?

Yogesh Kulkarni

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)

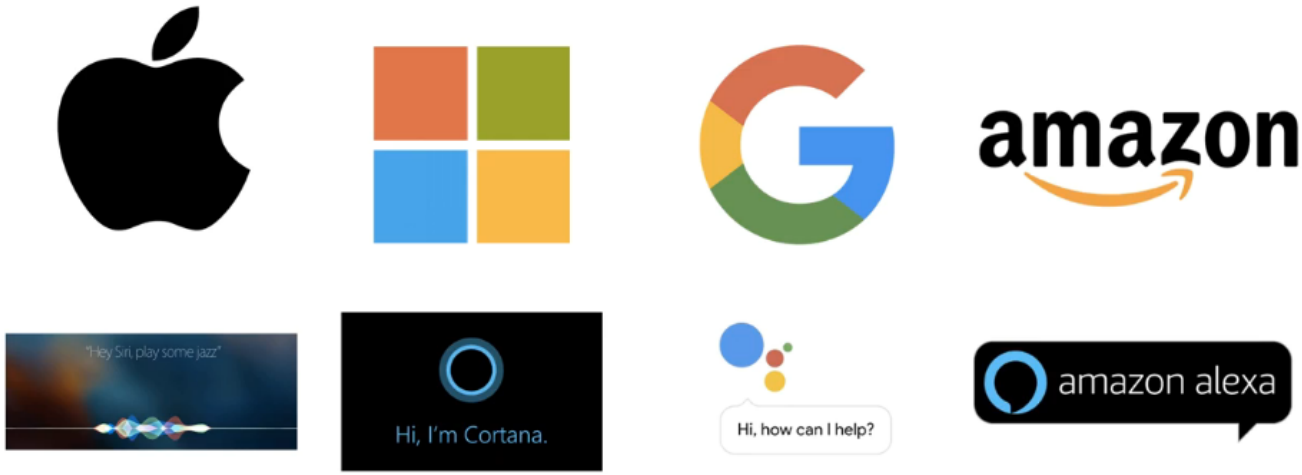
Instead, how about typing/saying your query directly and getting the answer right away?

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.

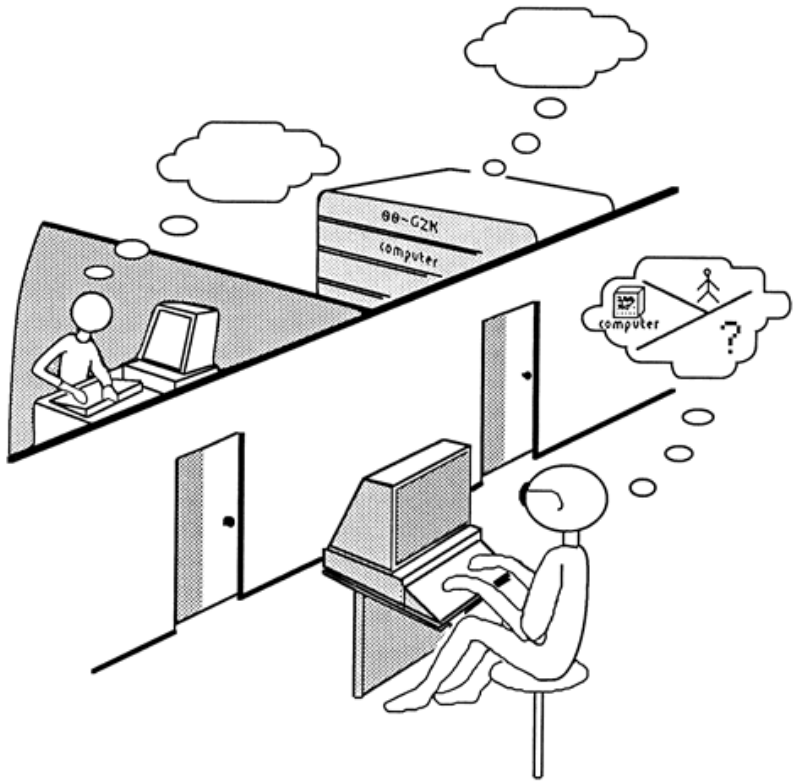
Its hard to master language, and thus NLP.

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

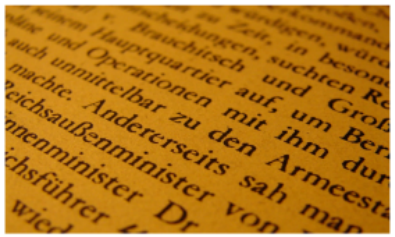
- In 1990, Hugh Loebner started Turing Test competition
- $100,000 will be awarded to the first bot that judges cannot distinguish from a real human in a Turing test that includes text, visual, and auditory input.
- Nobody has won the grand prize yet.
- 2016 (and 2013) year-wise top winner - Mitsuku.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

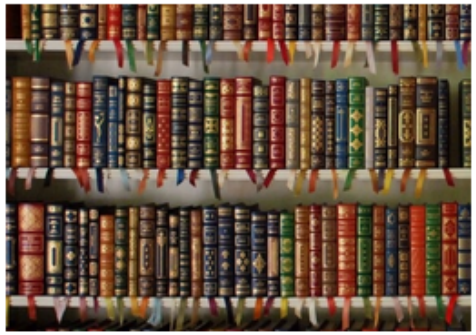
THIS WEEK

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)

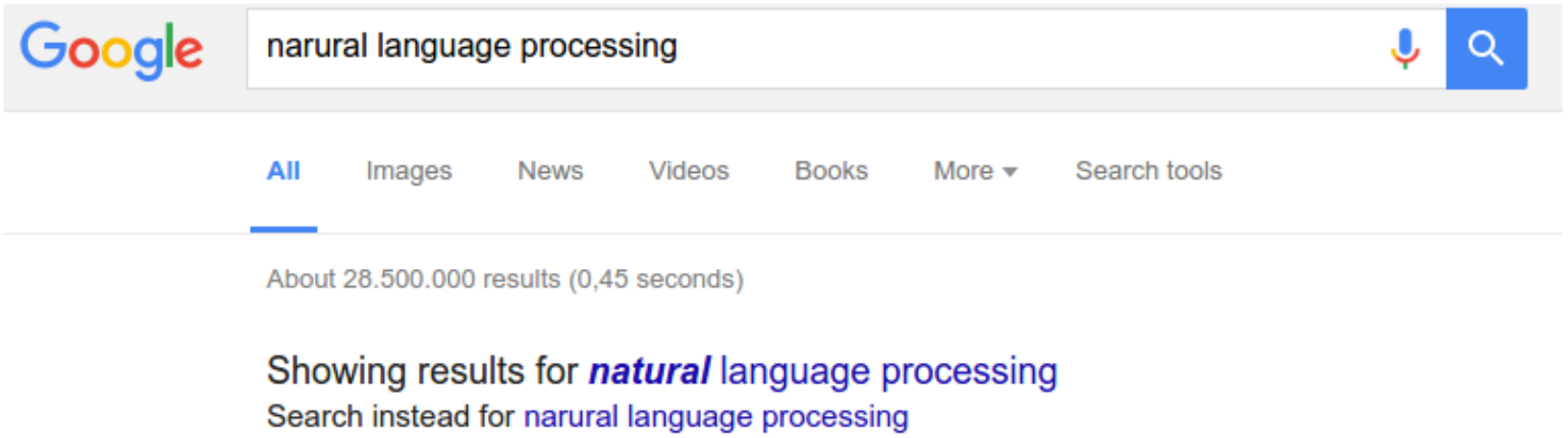
(http://www.businessnewsdaily.com/2023-celebrity-hiring.html)

NLP Applications

Grammar

Spell and Grammar Checking

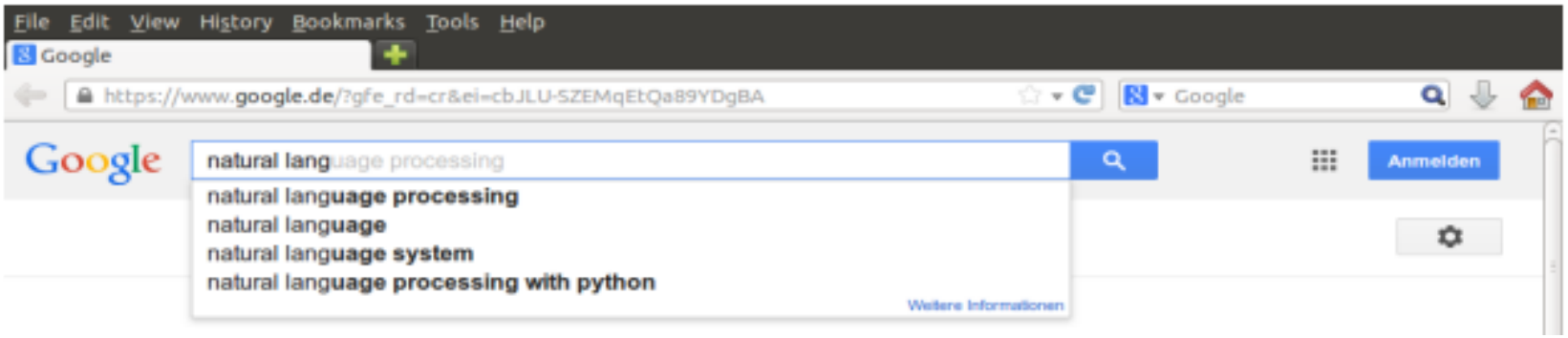
- Checking spelling and grammar
- Suggesting alternatives for the errors



Word Prediction

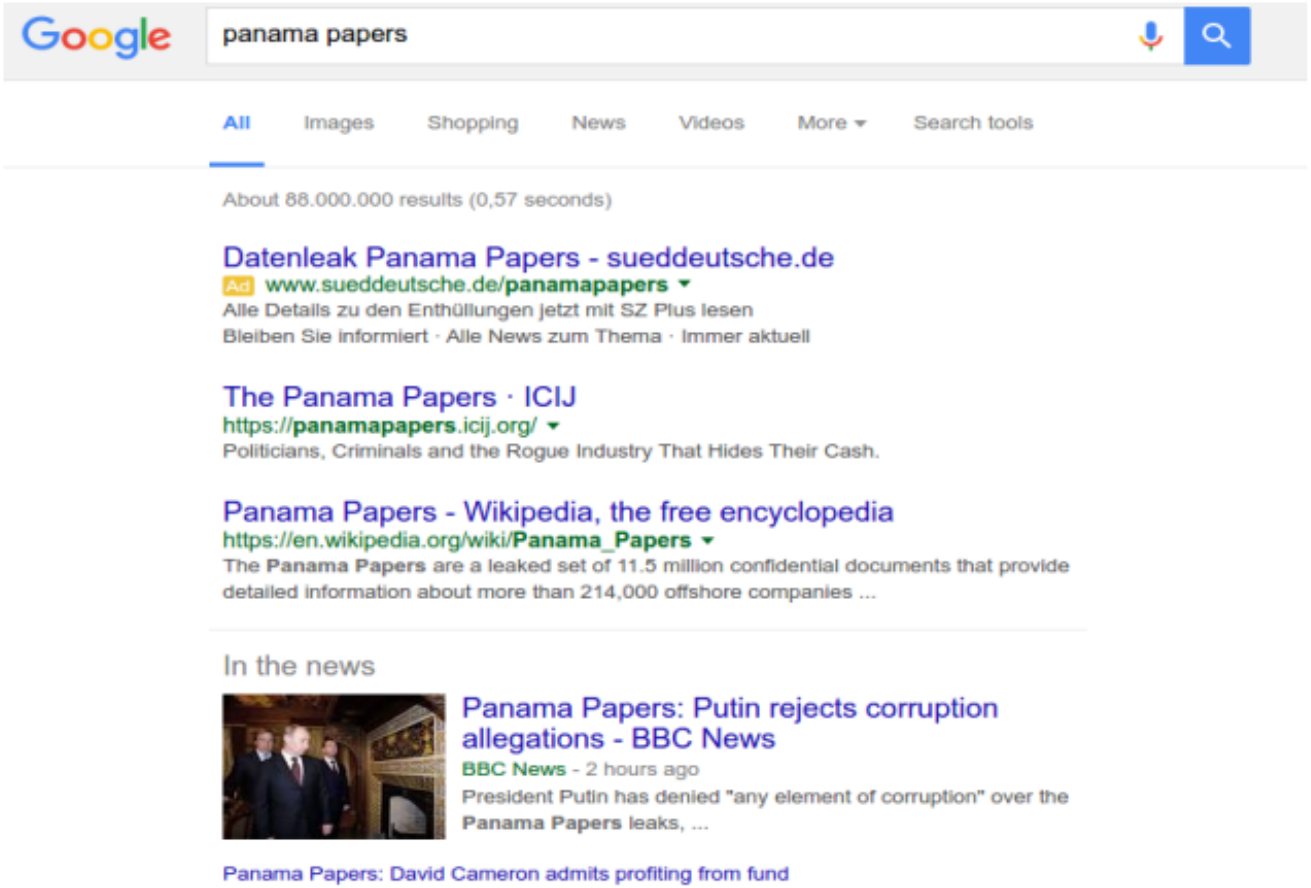
Word Prediction: Predicting the next word that is highly probable to be typed by the user

- Mobile typing
- Search Engines



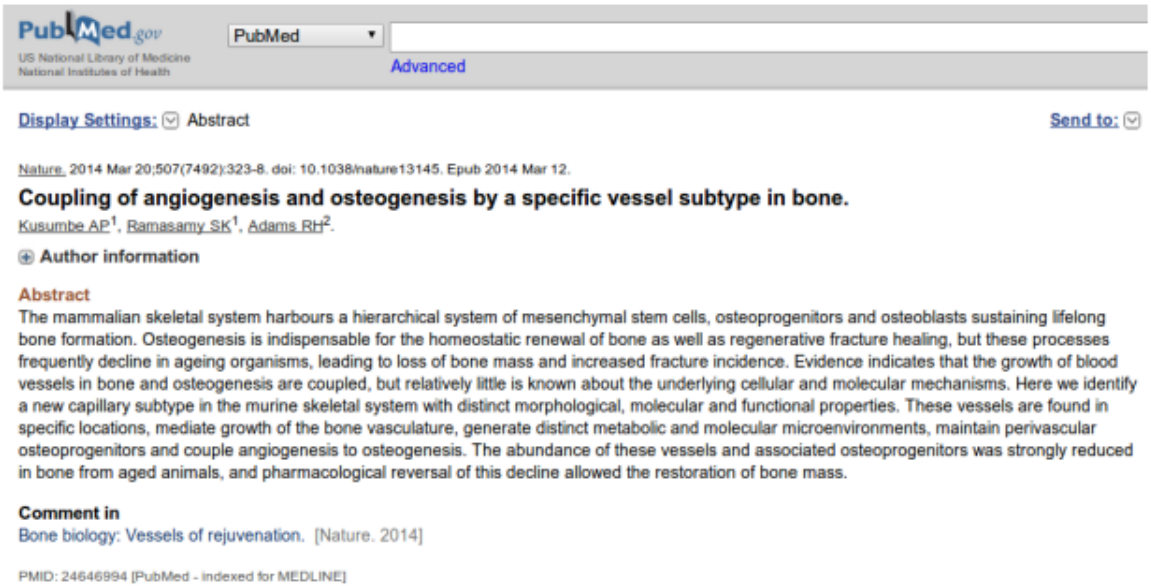
Information Retrieval

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]

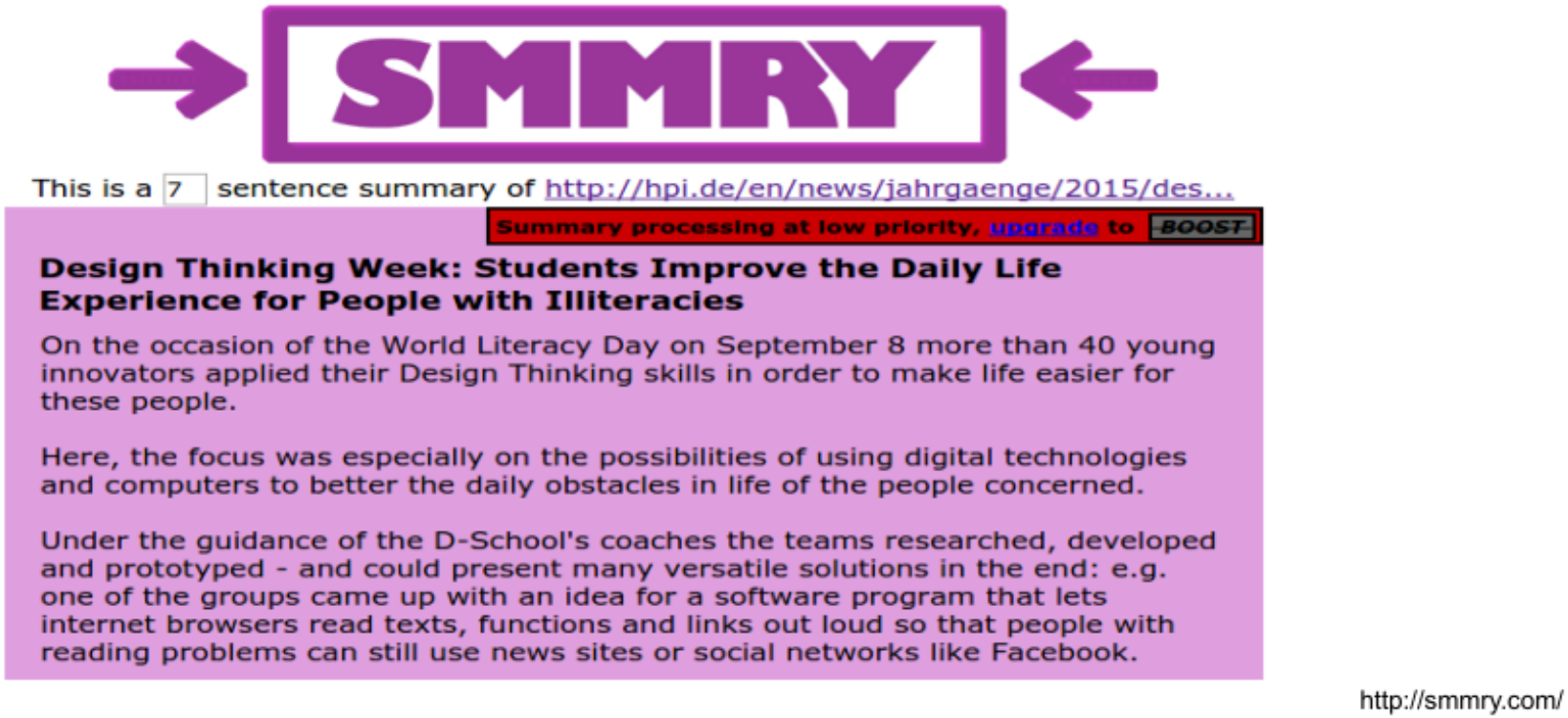
MeSH Terms

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

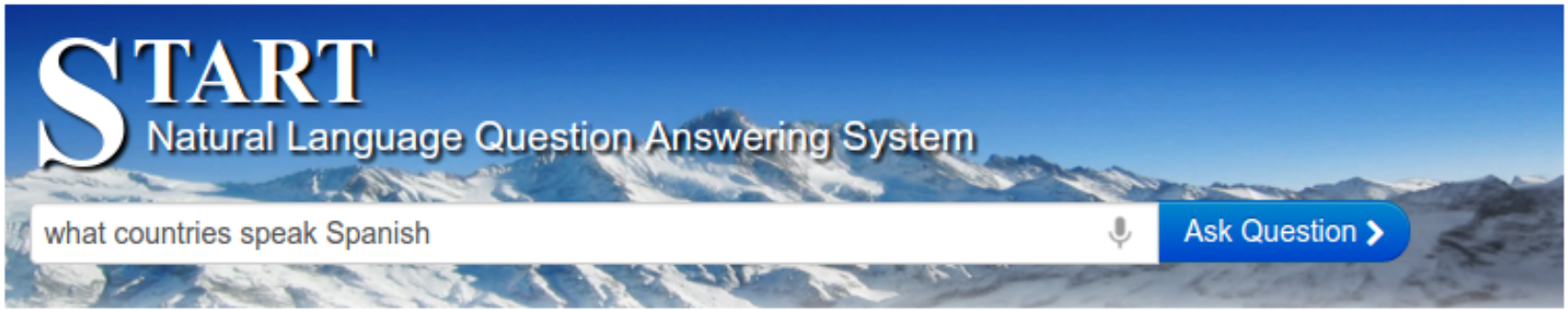
Summarization: Generating a short summary from one or more documents, sometimes based on a given query



http://smmry.com/

Question answering

Question answering: Answering questions with a short answer



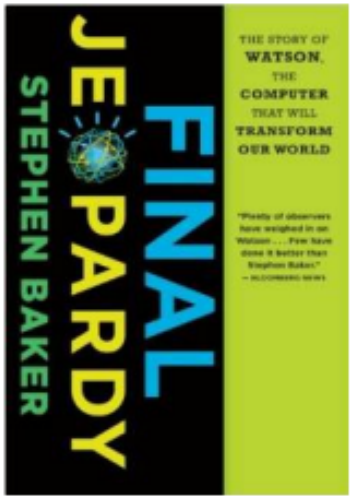
===> what countries speak Spanish

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

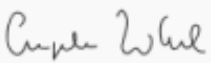
Chancellor	Helmut Kohl
Preceded by	Klaus Töpfer
Succeeded by	Jürgen Trittin

Minister for Women and Youth

In office
18 January 1991 – 17 November 1994

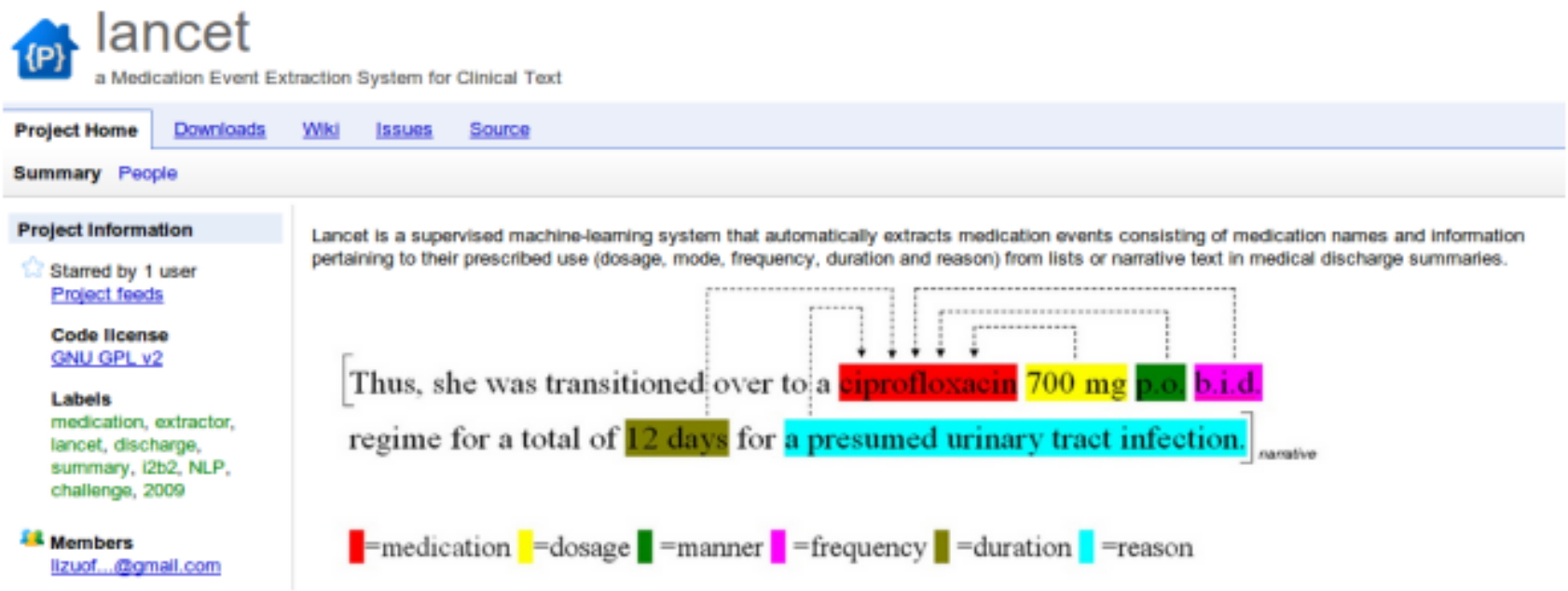
Chancellor	Helmut Kohl
Preceded by	Ursula Lehr
Succeeded by	Claudia Nolte

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...

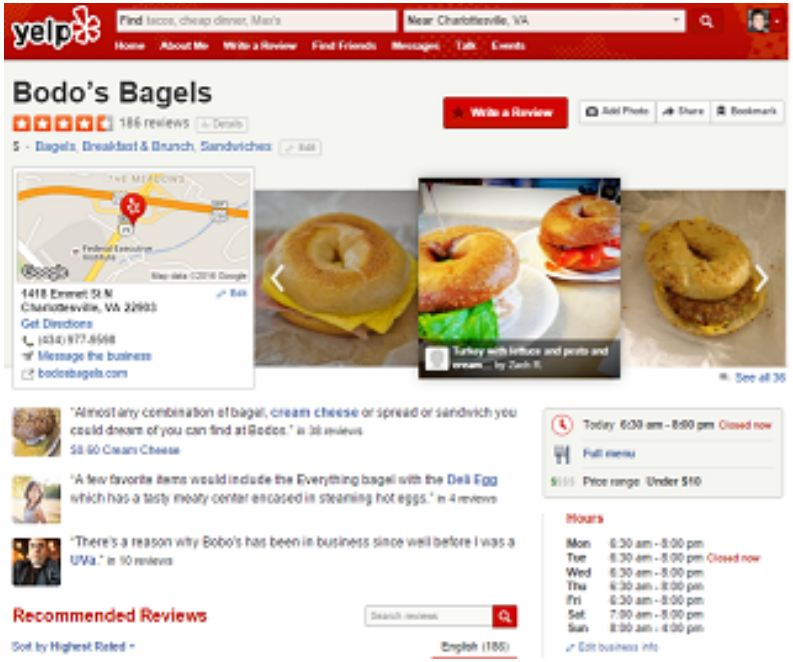
Read the full review ›

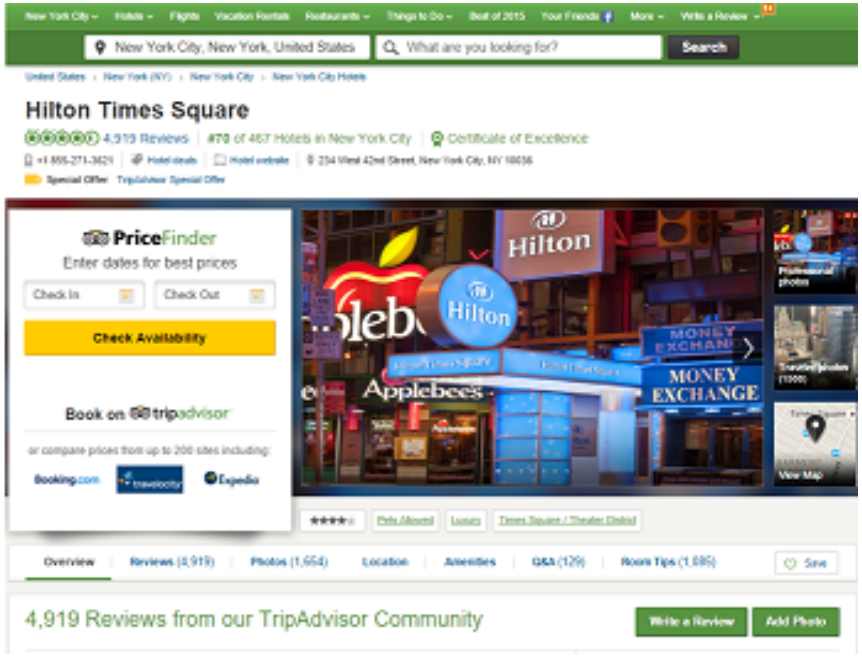
Published on April 2, 2009 by P. Nadkarni

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

Sentiment Analysis

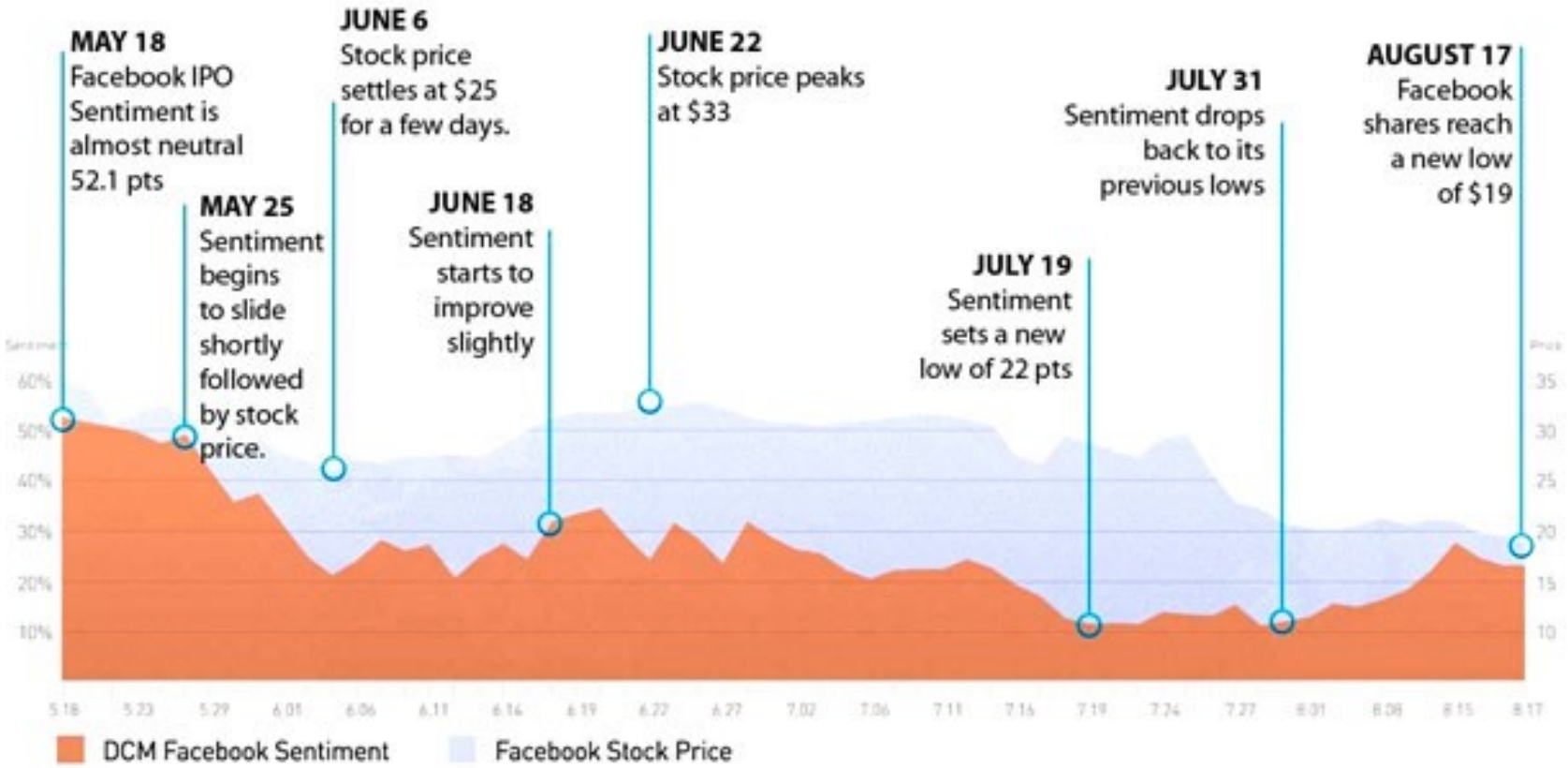
Restaurant/hotel recommendation, Product reviews





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

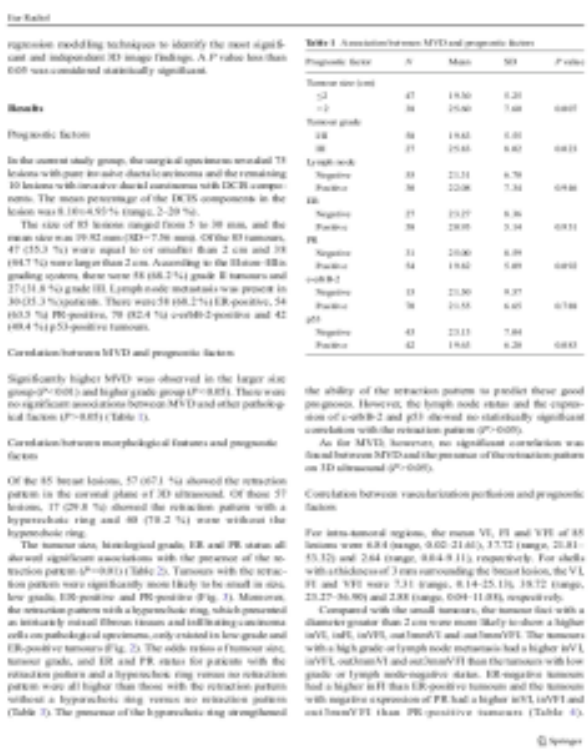
- Easy (mostly solved)
  - Spell and grammar checking
  - Some text categorization tasks
  - Some named-entity recognition tasks
- Intermediate (good progress)
  - Information retrieval
  - Sentiment analysis
  - Machine translation
  - Information extraction
- Difficult (still hard)
  - Question answering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

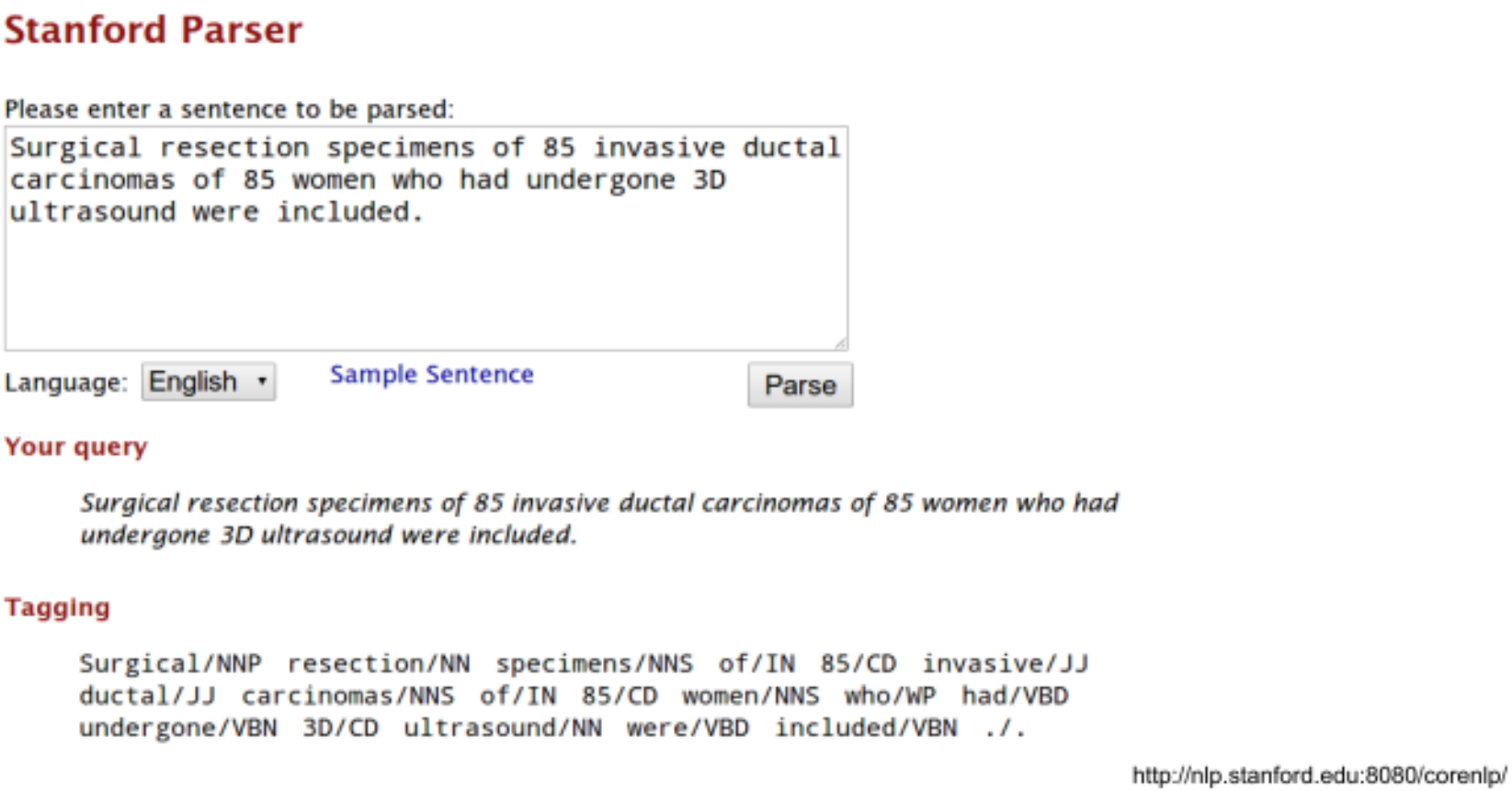
Lemmatization

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

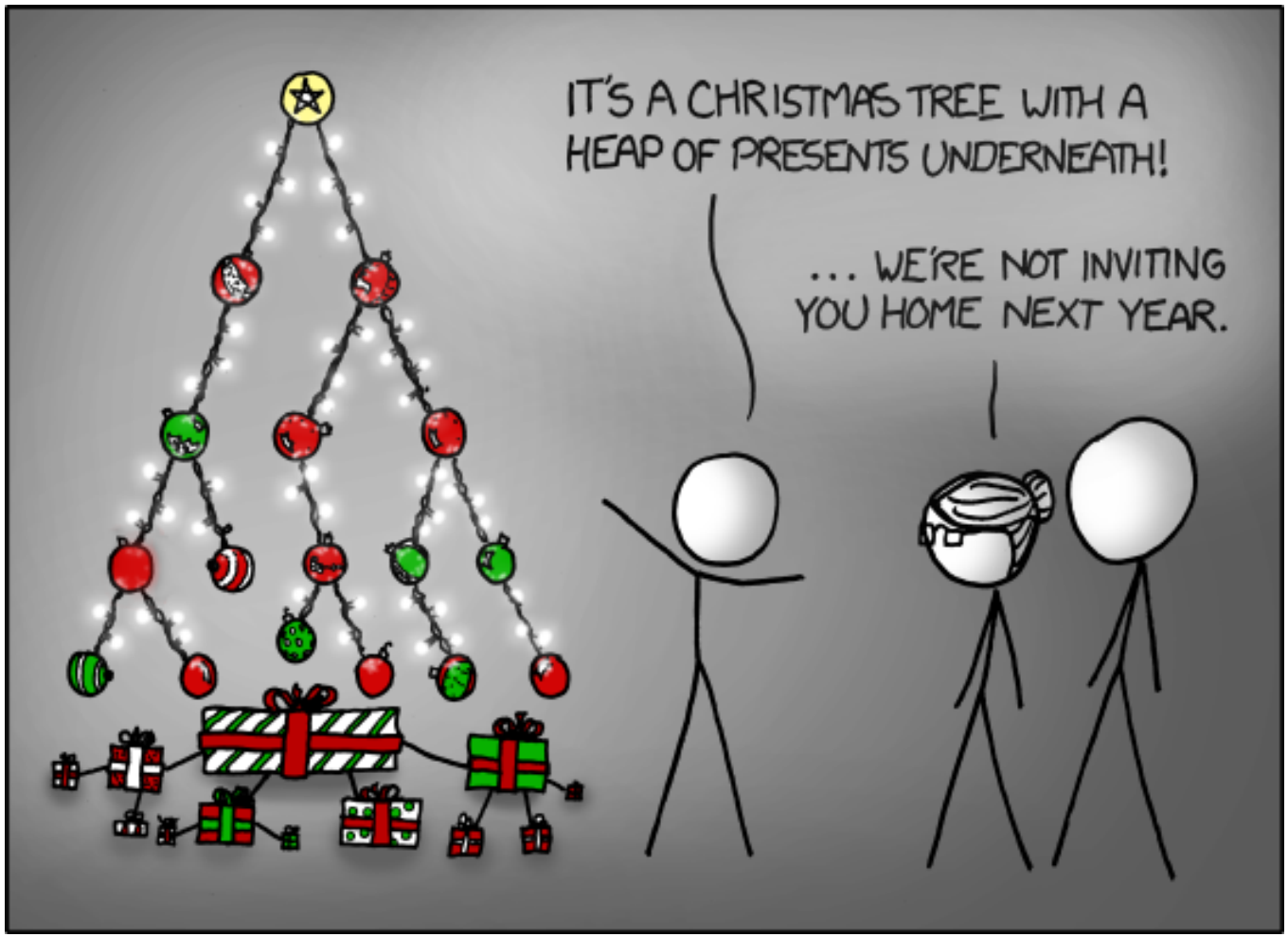
Parse

```
(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

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



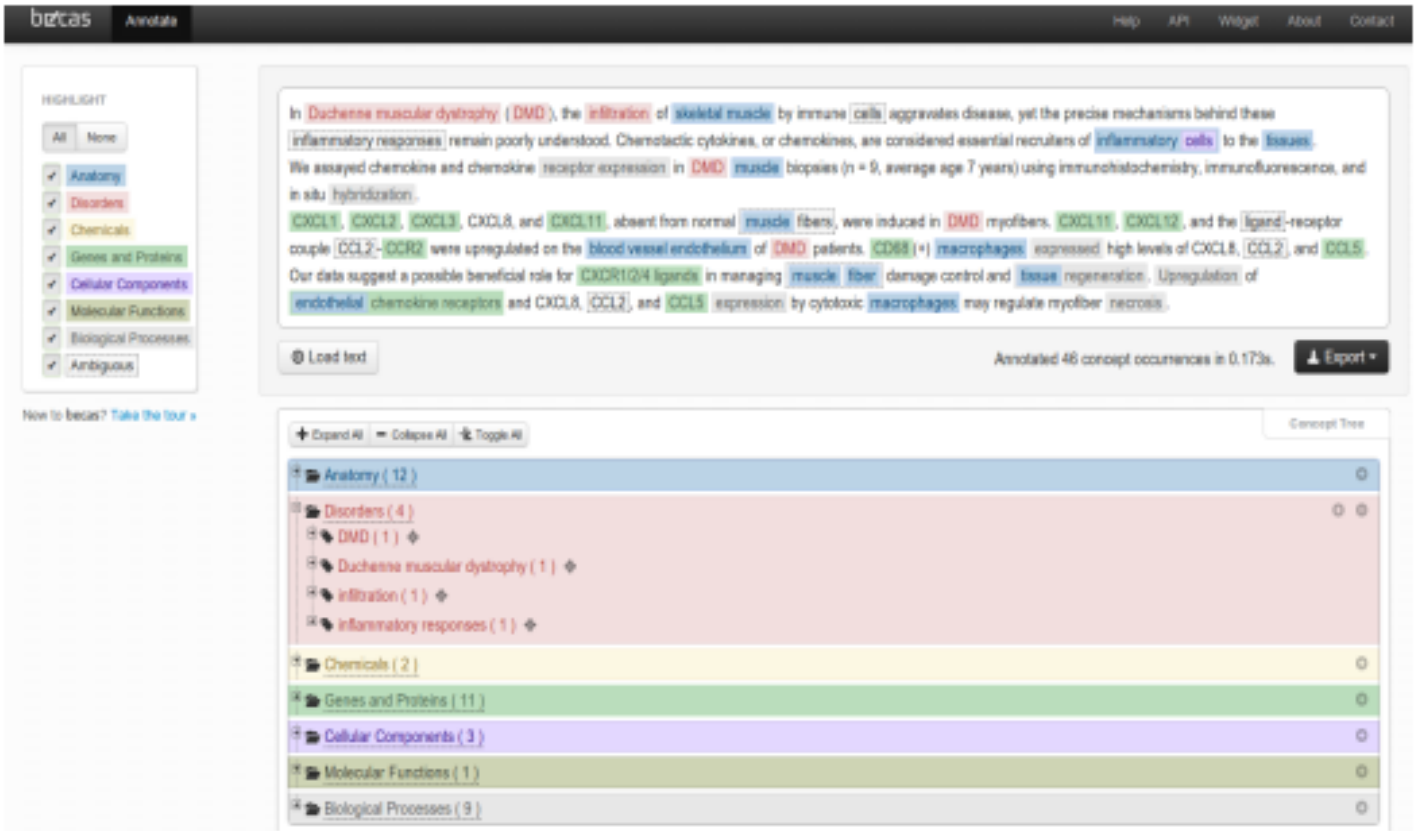
Syntax Tree

Syntax: Sample English grammar

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

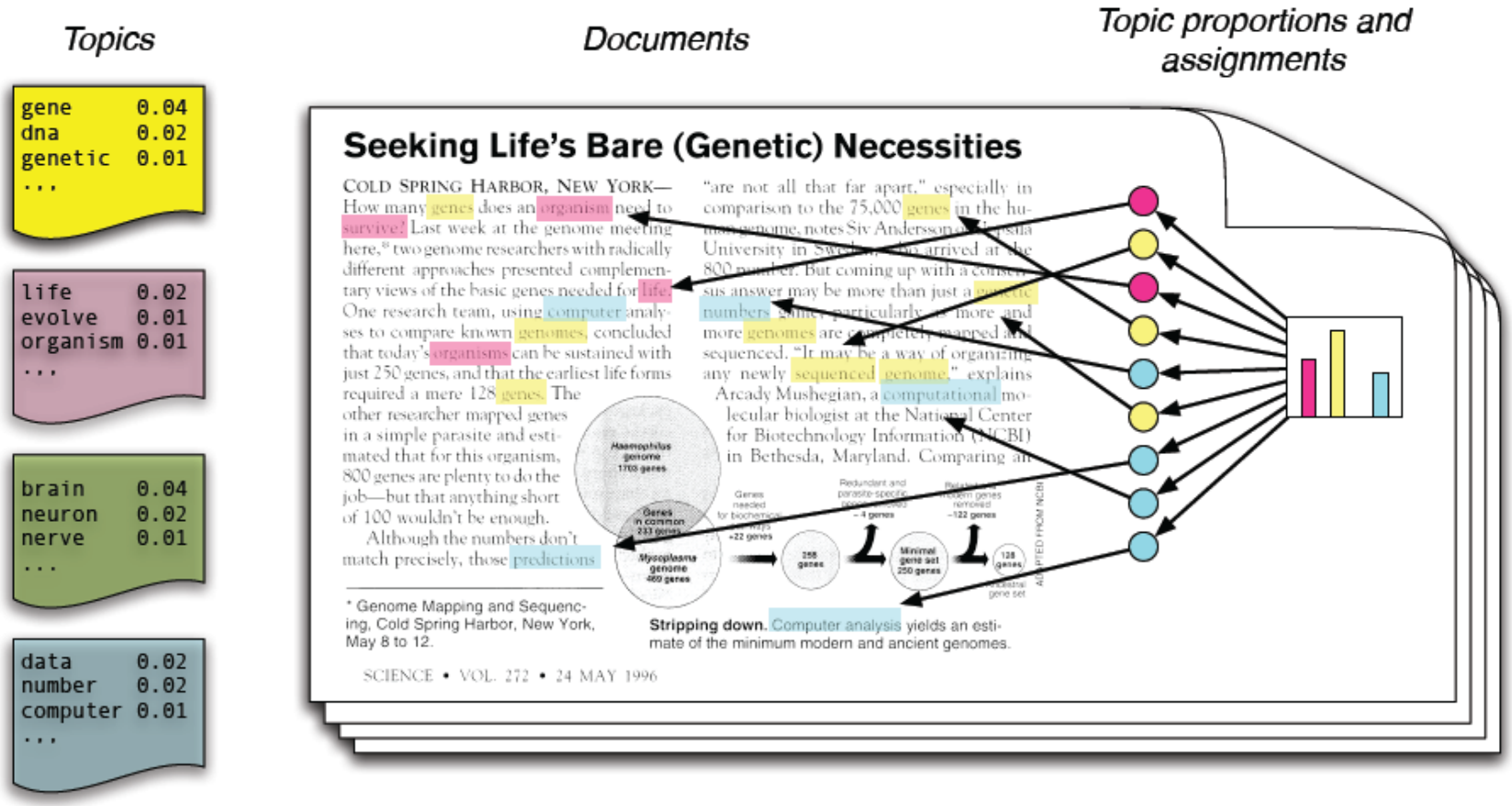
Named-entity recognition

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



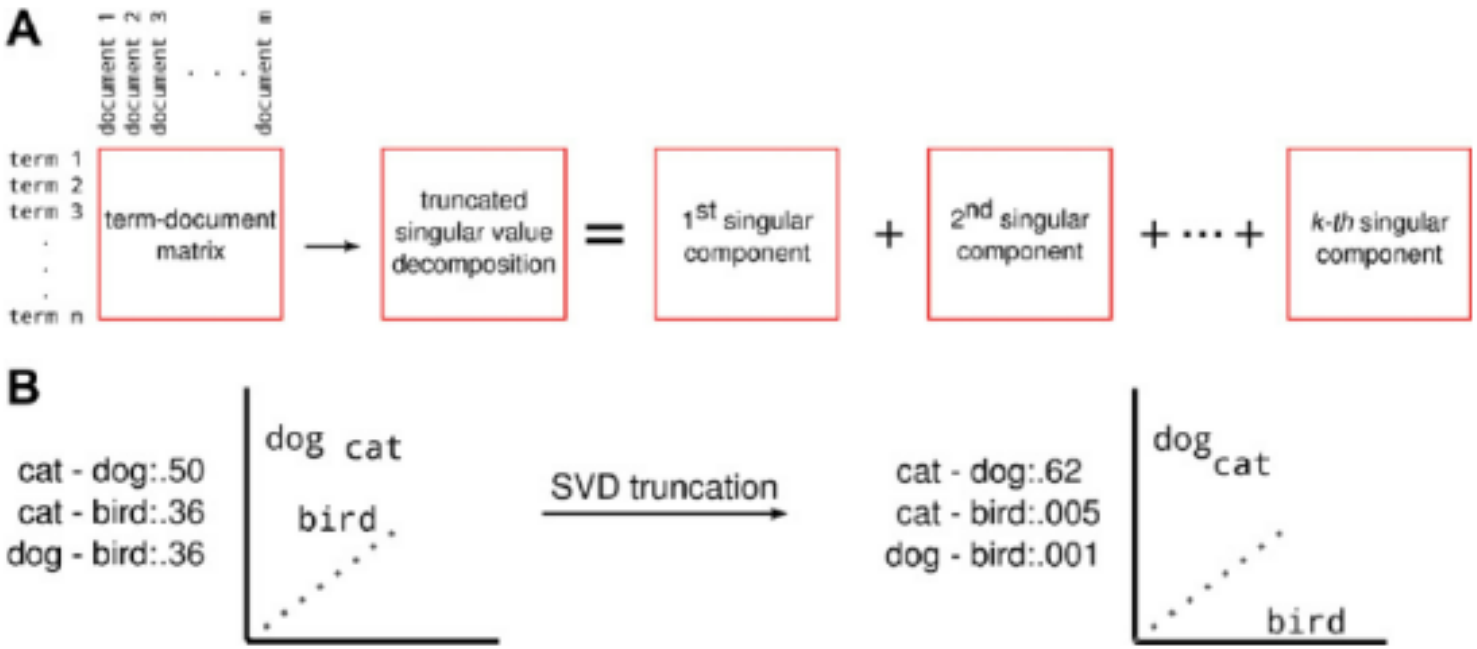
Topic modelings

Topic modeling: Identifying structures in the text corpus



Word embeddings

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



Language Resources and Text Pre- Processing

Text Resources

Text is available at multiple locations and in multiple formats:

- Web sites
- Databases
- Social Networks
- MS Word and Pdf files
- Etc.

The collection of text is called as "Corpora' (plural of 'Corpus').
There are often specialized APIs that let you access text from different platforms.
Not all text follows same representation (encoding) scheme.

Corpus

A corpus is a collection of documents to learn about.
(labeled or unlabeled)

Text Encoding

- ASCII is character encoding system that only supports only 128 different characters (for 7-bit encoding system or 255 for single byte):
  - Sufficient for English text (128 characters are enough!!!)
  - Insufficient for many other languages
  - How to accommodate ALL of them?
- Unicode supports over a million characters:
  - A single character set to cover all
  - Each character is assigned a number, called a code point
  - Code points are four hex numbers (`\uXXXX`)

Unicode: characters are not glyphs

- In Unicode, characters are abstract entities and can have more than one glyphs (written shapes)
- A font system and additional algorithms take care after proper representation of character codes
- UTF-8 (amongst other encodings) uses multiple bytes and can represent the full range of Unicode characters.

Exercise

- Create a UTF-8 file using a text editor
- Read the file using ASCII encoding
- Read the file using UTF-8 encoding
- Compare the outputs
- Convert the encoding of the file into Latin2

```
f = codecs.open(path,'w',encoding='latin2')
```

Corpora

- A text corpus is a large, structured collection of texts.
- The Open Language Archives Community (OLAC) provides an infrastructure for documenting and discovering language resource
- OLAC is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by:
  - developing consensus on best current practice for the digital archiving of language resources, and
  - developing a network of inter-operating repositories and services for housing and accessing such resources.
- http://www.language-archives.org/
- NLTK comes with many corpora

Annotated Text Corpora

- Many text corpora contain linguistic annotations, representing genres, POS tags, named entities, syntactic structures, semantic roles, and so forth.
- Not part of the text in the file; it explains something of the structure and/or semantics of text
- Grammar annotation
- Semantic annotation
- Lower level annotation
  - Word tokenization
  - Sentence Segmentation
  - Paragraph Segmentation

Text Corpus Structure

- The simplest kind lacks any structure (i.e annotation): it is just a collection of texts (Gutenberg, web text)
- Often, texts are grouped into categories that might correspond to genre, source, author, language, etc. (Brown)
- Sometimes these categories overlap, notably in the case of topical categories as a text can be relevant to more than one topic. (Reuters)
- Occasionally, text collections have temporal structure (news collections, Inaugural Address Corpus)

Brown Corpus

The Brown Corpus was the first million-word electronic corpus of English, created in 1961 at Brown University. This corpus contains text from 500 sources, and the sources have been categorized by genre, such as news, editorial, and so on.

ID	File	Genre	Description
A16	ca16	news	Chicago Tribune: *Society Reportage*
B02	cb02	editorial	Christian Science Monitor: *Editorials*
C17	cc17	reviews	Time Magazine: *Reviews*
D12	cd12	religion	Underwood: *Probing the Ethics of Realtors*
E36	ce36	hobbies	Norling: *Renting a Car in Europe*
F25	cf25	lore	Boroff: *Jewish Teenage Culture*
G22	cg22	belles_lettres	Reiner: *Coping with Runaway Technology*
H15	ch15	government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
J17	cj19	learned	Mosteller: *Probability with Statistical Applications*
K04	ck04	fiction	W.E.B. Du Bois: *Worlds of Color*
L13	cl13	mystery	Hitchens: *Footsteps in the Night*
M01	cm01	science_fiction	Heinlein: *Stranger in a Strange Land*
N14	cn15	adventure	Field: *Rattlesnake Ridge*
P12	cp12	romance	Callaghan: *A Passion in Rome*
R06	cr06	humor	Thurber: *The Future, If Any, of Comedy*

Brown Corpus

- The Brown Corpus is a convenient resource for studying systematic differences between genres, a kind of linguistic inquiry known as stylistics.
- For example, we can compare genres in their usage of modal verbs:

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

Reuters Corpus

- The Reuters Corpus contains 10,788 news documents totaling 1.3 million words.
- The documents have been classified into 90 topics, and grouped into two sets, called “training“ and “test”
- This split is for training and testing algorithms that automatically detect the topic of a document
- Unlike the Brown Corpus, categories in the Reuters corpus overlap with each other, simply because a news story often covers multiple topics.

Lexical Resources

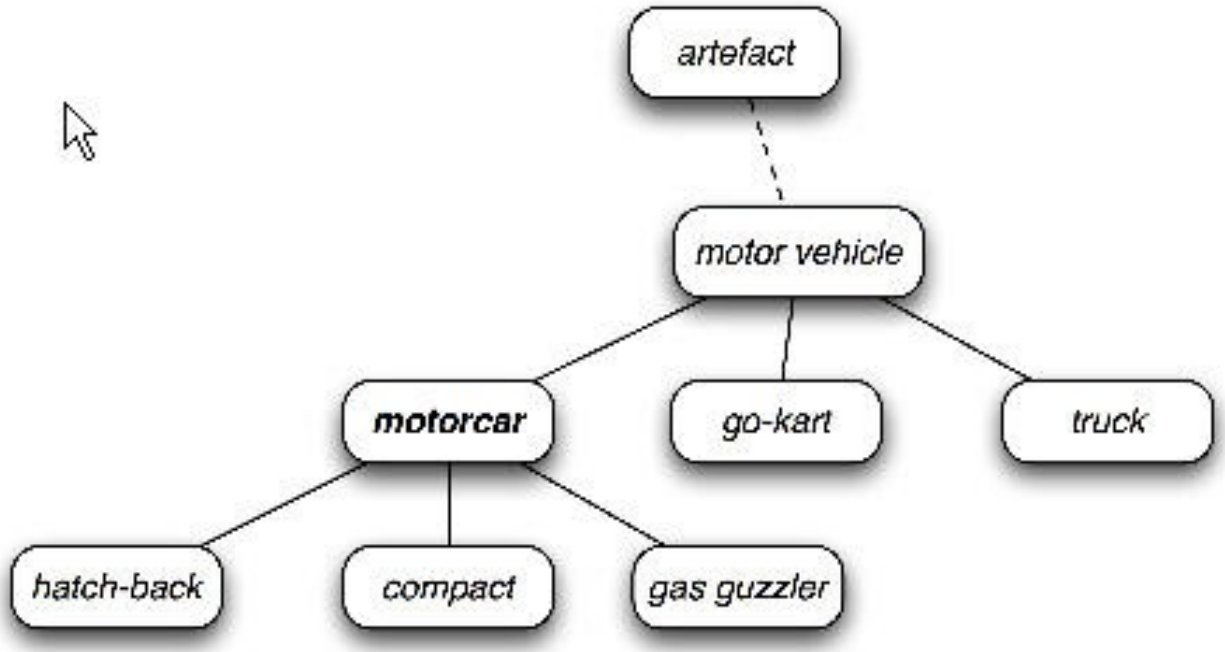
- A lexicon, or lexical resource, is a collection of words and/or phrases along with associated information such as part of speech and sense definitions.
- Lexical resources are secondary to texts, and are usually created and enriched with the help of texts
- A vocabulary (list of words in a text) is the simplest lexical resource
- Lexical entry
- A lexical entry consists of a headword (also known as a lemma) along with additional information such as the part of speech and the sense definition.
- WordNet
- VerbNet
- FrameNet

WordNet

- WorldNet is a semantically-oriented dictionary of English, similar to a traditional thesaurus but with a richer structure.
- WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept*.
- Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser.
- WordNet is also freely and publicly available for download.
- WordNet's structure makes it a useful tool for computational linguistics and natural language processing.
- NLTK includes the English WordNet, with 155,287 words and 117,659 synonym sets

The WordNet Hierarchy

- It's very easy to navigate between concepts. For example, given a concept like motorcar, we can look at the concepts that are more specific; the (immediate) hyponyms.



WordNet: Semantic Similarity

- Knowing which words are semantically related is useful for indexing a collection of texts, so that a search for a general term like vehicle will match documents containing specific terms like limousine.
- Two synsets linked to the same root may have several hypernyms in common. If two synsets share a very specific hypernym - one that is low down in the hypernym hierarchy - they must be closely related.

VerbNet: A Verb Lexicon

- VerbNet, a hierarchical verb lexicon linked to WordNet. It can be accessed with nltk.corpus.verbnet.
- *VerbNet is the largest on-line verb lexicon currently available for English.
- It is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to other lexical resources such as WordNet and FrameNet.

FrameNet

- Project housed at the International Computer Science Institute (ICSI) in Berkeley, California which produces an electronic resource based on semantic frames. http://framenet.icsi.berkeley.edu/
- 11,600 lexical units, in more than 960 semantic frames, exemplified in more than 150,000 annotated sentences.

FrameNet

Travel

Definition:

In this frame a Traveler goes on a journey, an activity, generally planned in advance, in which the Traveler moves from a Source location to a Goal along a Path or within an Area. The journey can be accompanied by Co_participants and Baggage. The Duration or Distance of the journey, both generally long, may also be described as may be the Mode_of_transportation. Words in this frame emphasize the whole process of getting from one place to another, rather than profiling merely the beginning or the end of the journey.

Ellen JOURNEYED to Europe with five suitcases.

Samantha JOURNEYED 2500 miles with her family by sea to China.

The Osbournes took a TRIP from Beverly Hills to London on the Concorde.

FEs:

Core:

Area [Area]
Semantic Type
Location
This is the Area in which the traveling takes place. This frame element describes the enclosed area inside which travelling, of unspecified Source, Path or Goal takes place.

We TRAVELLED in Europe.

Direction [dir]
The direction in which the Traveler goes.

They began their ODYSSEY north.

Goal [Goal]
Semantic Type
Goal
The Goal is the location where the travelers end up.

Mode_of_transportation [MoT]
The Mode_of_transportation expresses how the motion of the Traveler is effected, by their body or by a vehicle which holds and conveys the Traveler. Vehicles can move in any way and in any medium. They are usually expressed obliquely with 'in' or 'by'.

Barney used to TRAVEL by bus a lot

Strom TRAVELLED on foot to see the Pope

Path [Path]
Semantic Type
Path
The Path is the route along which the travel takes place.

Source [Src]
Semantic Type
Source
The Source is the starting point of the trip.

Text Normalization

- Stemming
- Convert to lower case
- Identifying non-standard words including numbers, abbreviations, and dates, and mapping any such tokens to a special vocabulary.
- For example, every decimal number could be mapped to a single token 0.0, and every acronym could be mapped to AAA. This keeps the vocabulary small and improves the accuracy of many language modeling tasks.
- Lemmatization
- Make sure that the resulting form is a known word in a dictionary
- WordNet lemmatizer only removes affixes if the resulting word is in its dictionary

The NLP Pipeline (Recap)

For a given problem to be tackled:

- Choose corpus (or build your own)
- Choose annotation to use (or choose the label set and label it yourself)
- Choose or implement new NLP algorithms

Classification

Question

Why do we need to classify texts?

Why do we need to classify texts?

- As a self-sufficient task:
  - Spam filtering
  - Sentiment analysis
  - Fake news/clickbait detection
  - Troll/bot protection
- As a part of more complicated NLP tasks
  - Data filtering
  - Intent classification in dialog systems
  - Hybrid machine translation systems :-)

(Ref: Text Classification - Elena Voita, Yandex Research)

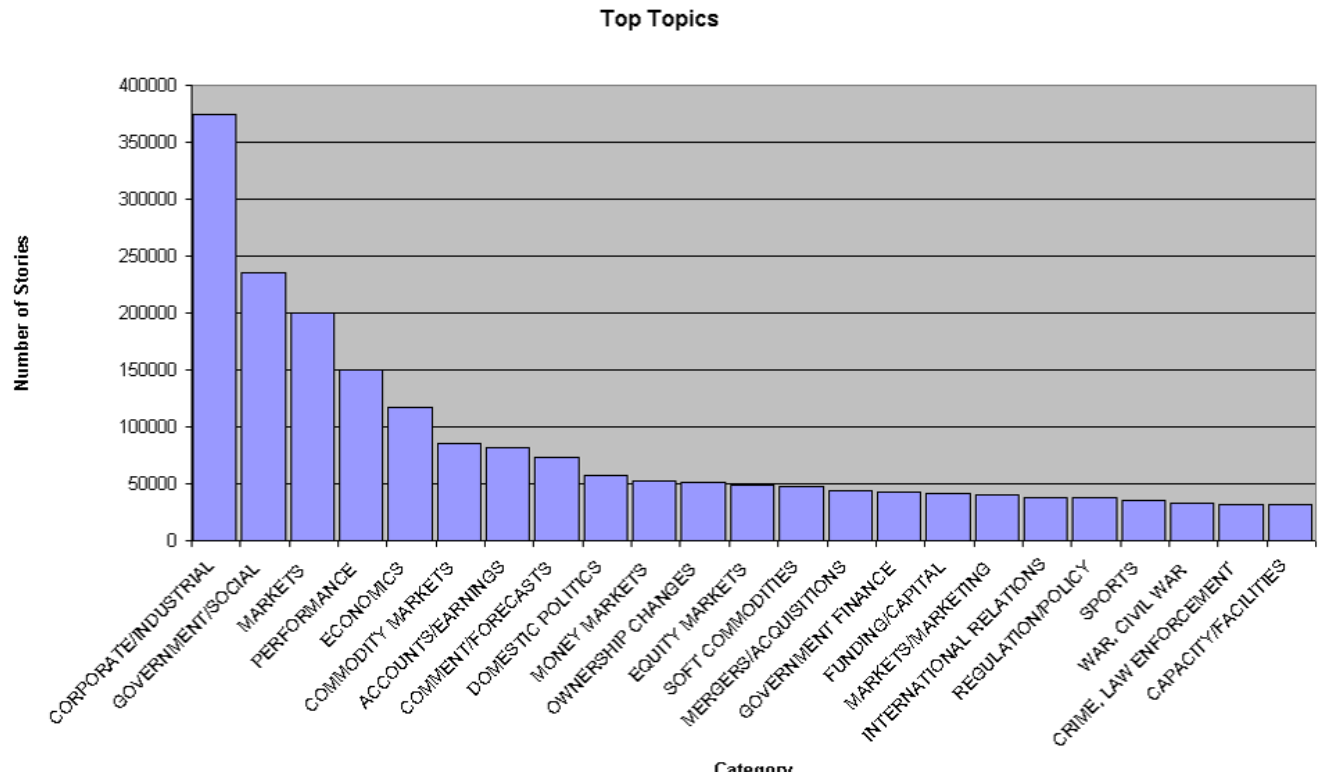
Text classification Applications

- Web pages organized into category hierarchies
- Journal articles indexed by subject categories (e.g., the Library of Congress, MEDLINE, etc.)
- E-mail message filtering: Spam vs. anti-palm
- News events tracked and filtered by topics

News topic classification , example

Training Data: Reuters

- Gold standard
- Collection of (21,578) newswire documents.
- For research purposes: a standard text collection to compare systems and algorithms
- 135 valid topics categories



Reuters Sample

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off

tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

How Does Text Classification Work?

Manual and Automatic

- Manual: a human annotator interprets the content of text and categorizes it accordingly. This method usually can provide quality results but it's time-consuming and expensive.
- Automatic: applies machine learning, natural language processing, and other techniques to automatically classify text in a faster and more cost-effective way, but could be error-prone

Within Automatic

Three different types of systems:

- Rule-based systems
- Machine Learning based systems
- Hybrid systems

Classification Aspects

- Various issues regarding classification: Clustering vs. classification, binary vs. multi-way, flat vs. hierarchical classification, variants...
- Introduce the steps necessary for a classification task
  - Define classes (aka labels)
  - Label text
  - Define and extract features
  - Training and evaluation

What is Text Classification?

Assign the correct class label for a given input/object In basic classification tasks, each input is considered in isolation from all other inputs, and the set of labels is defined in advance.

Problem	Object	Label's categories
Tagging	Word	POS
Sense Disambiguation	Word	The word's senses
Information retrieval	Document	Relevant/not relevant
Sentiment classification	Document	Positive/negative
Text categorization	Document	Topics/classes
Author identification	Document	Authors
Language identification	Document	Language

Text Classification

Classify the document into semantics topics

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan defeated Belarus's Max Mirnyi and Vladimir Voltchkov to give the Americans an unsurmountable 3-0 lead in the best-of-five semi-final tie.

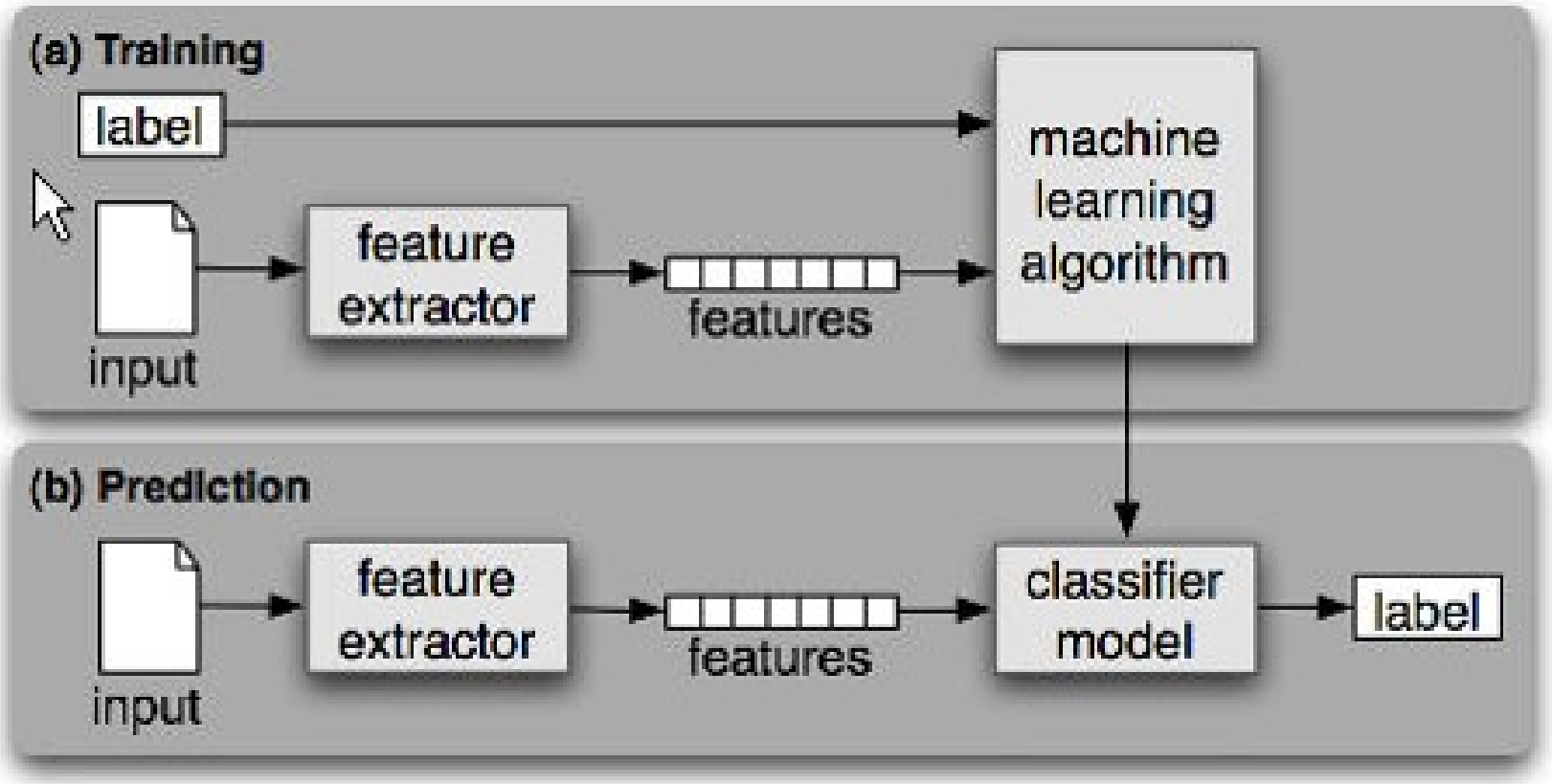
One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to the historic city of St. Augustine.

Classification vs. Clustering

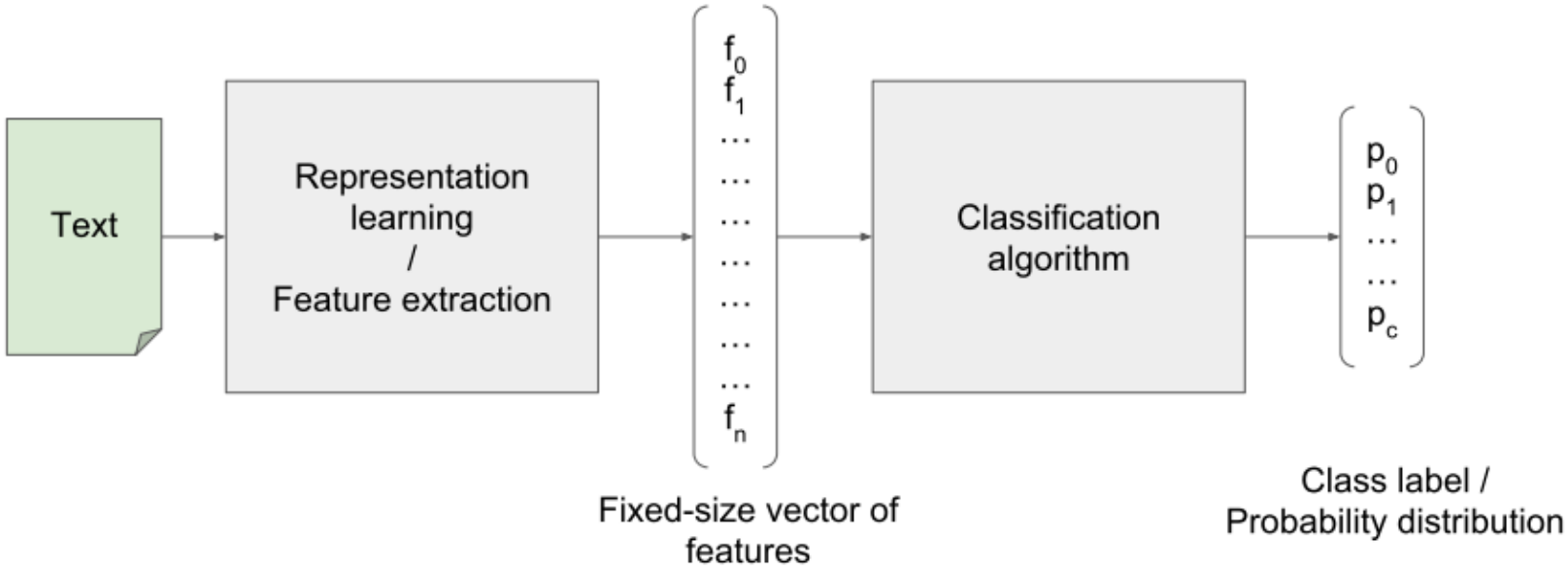
- Classification assumes labeled data: we know how many classes there are and we have examples for each class (labeled data).
- Classification is supervised
- In Clustering we don't have labeled data; we just assume that there is a natural division in the data and we may not know how many divisions (clusters) there are
- Clustering is unsupervised

Supervised classification

A classifier is called supervised if it is built based on training corpora containing the correct label for each input.



Text classification in general



(Ref: Text Classification - Elena Voita, Yandex Research)

Labels

- Binary classification: spam filtering/sentiment analysis
- Multi-class classification: categorization of goods
- Multi-label classification: #hashtag prediction

Training

- Adaptation of the classifier to the data
- Usually the classifier is defined by a set of parameters
- Training is the procedure for finding a "good" set of parameters
- Goodness is determined by an optimization criterion such as misclassification rate
- Some classifiers are guaranteed to find the optimal set of parameters

Algorithms for Classification

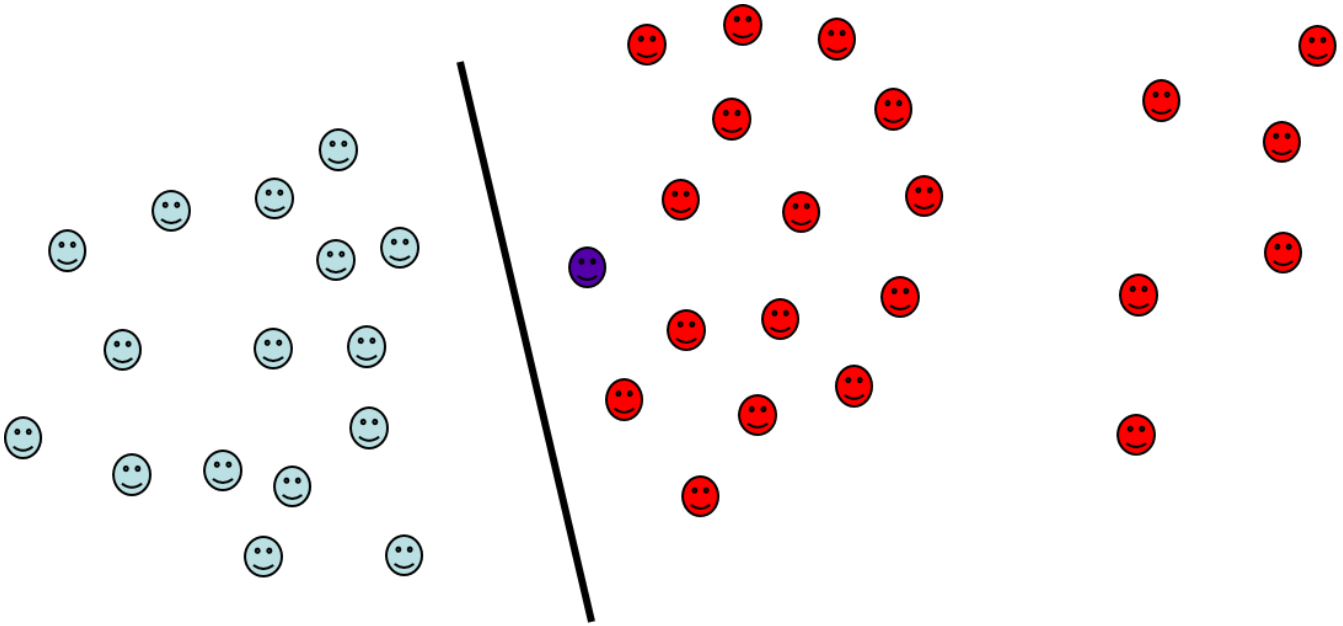
- It's possible to treat these learning methods as black boxes.
- But there's a lot to be learned from taking a closer look
- An understanding of these methods can help guide our choice for the appropriate learning method (binary or multi-class for example)
- Binary classification: Linear and not linear
- Multi-Class classification: Linear and not linear

Methods

- Linear Models:
  - Perceptron & Winnow (neural networks)
  - Large margin classifier
  - Support Vector Machine (SVM)
- Probabilistic models:
  - Naive Bayes
  - Maximum Entropy Models
- Decision Models: Decision Trees
- Instance-based methods: Nearest neighbor

(Linear) Classification

Choose the classier with the lower rate of misclassification



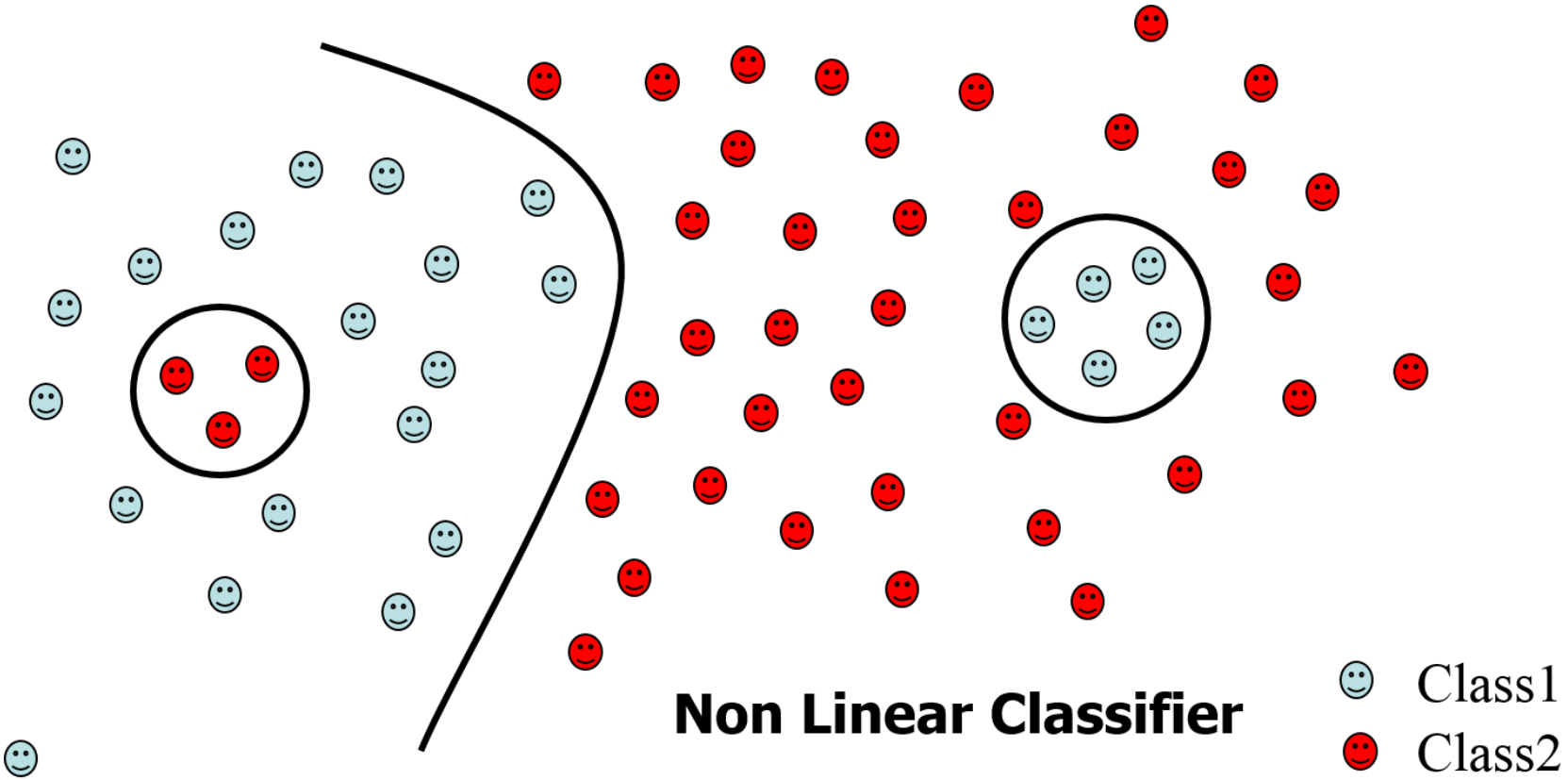
Linear classifier: $g(x) = \mathbf{w}x + \mathbf{w_0}$

For each set of parameters: $\mathbf{w}$, $\mathbf{w_0}$, calculate error

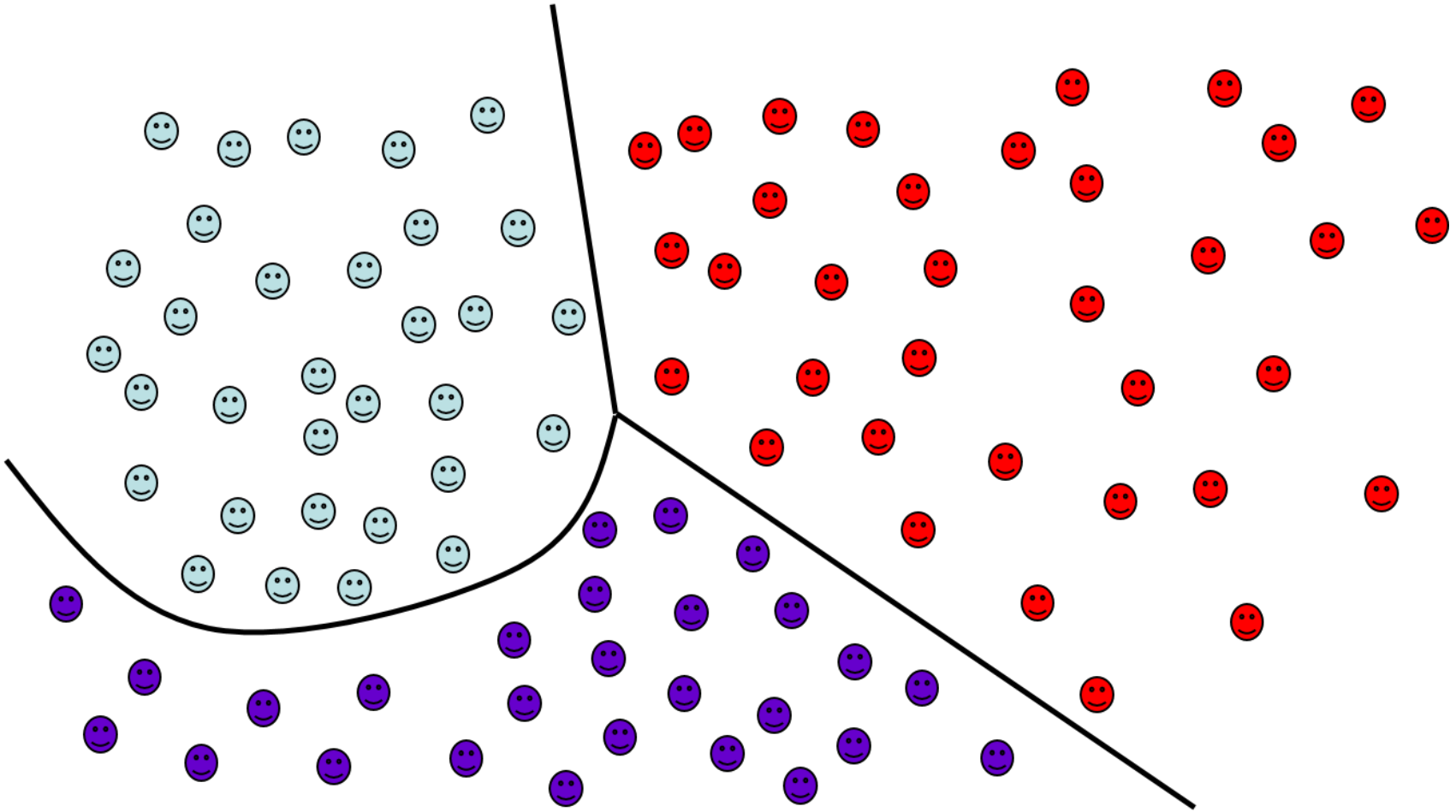
Class1

Class2

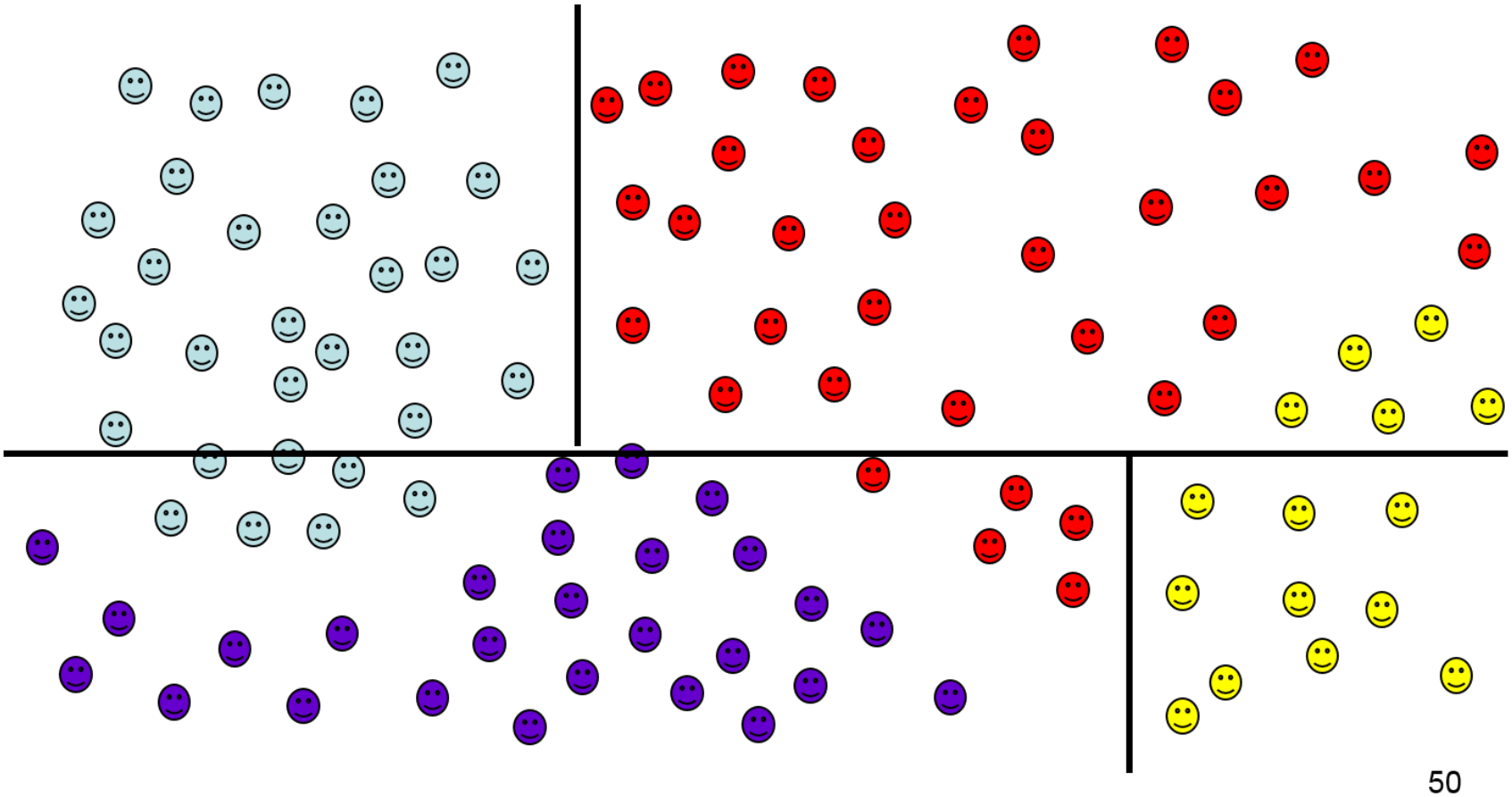
Non linearly separable data



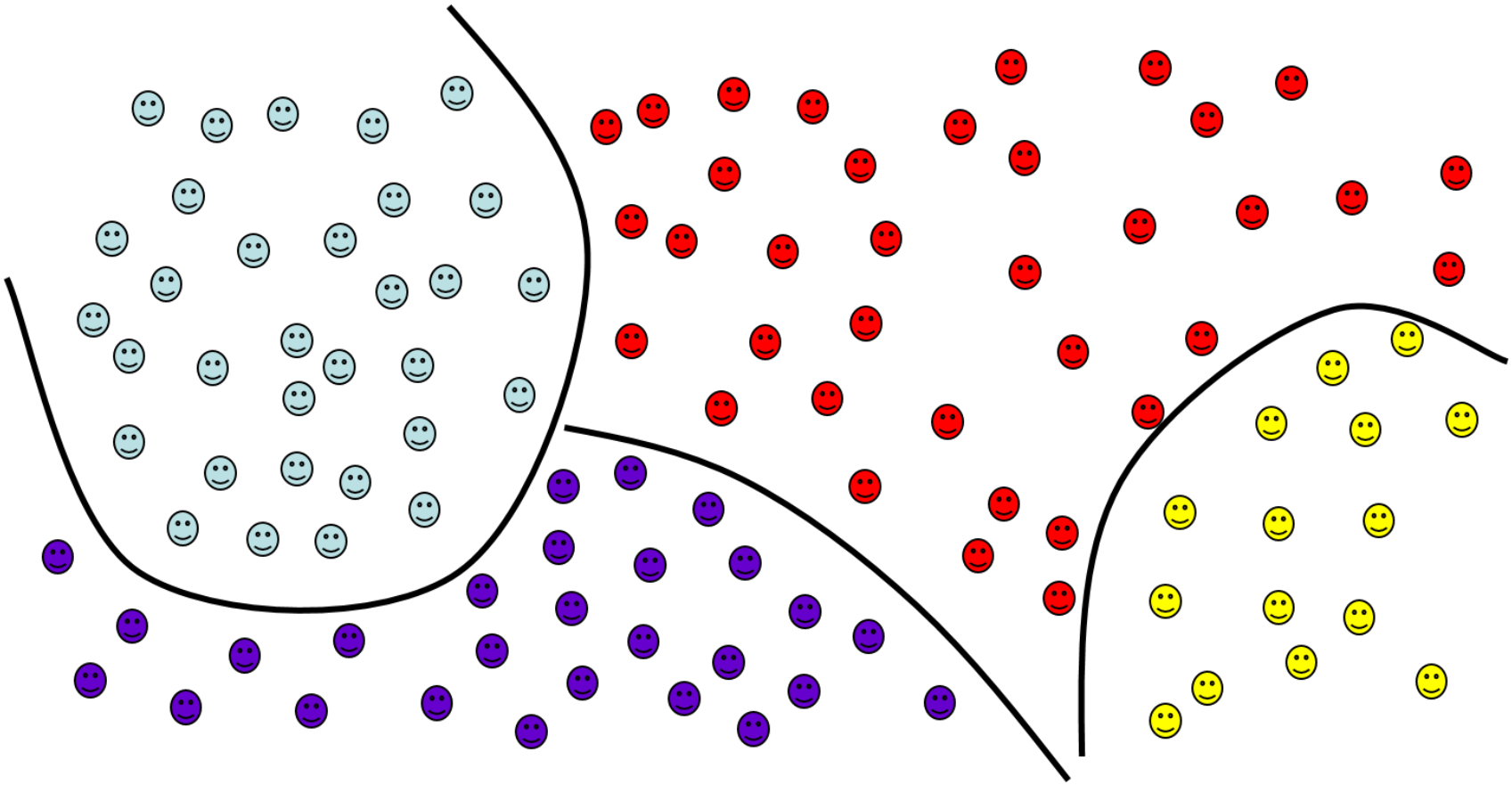
Multi-class classification



Linear, parallel class separators (ex: Decision Trees)

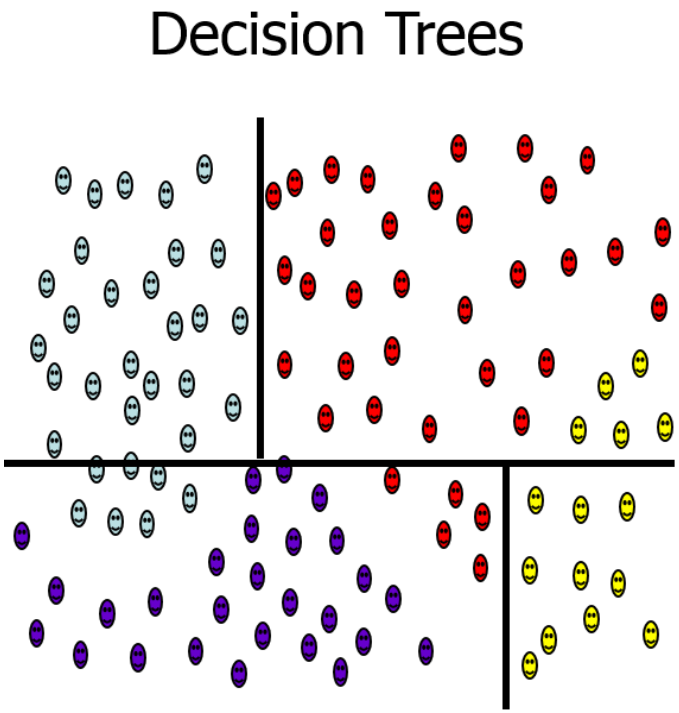


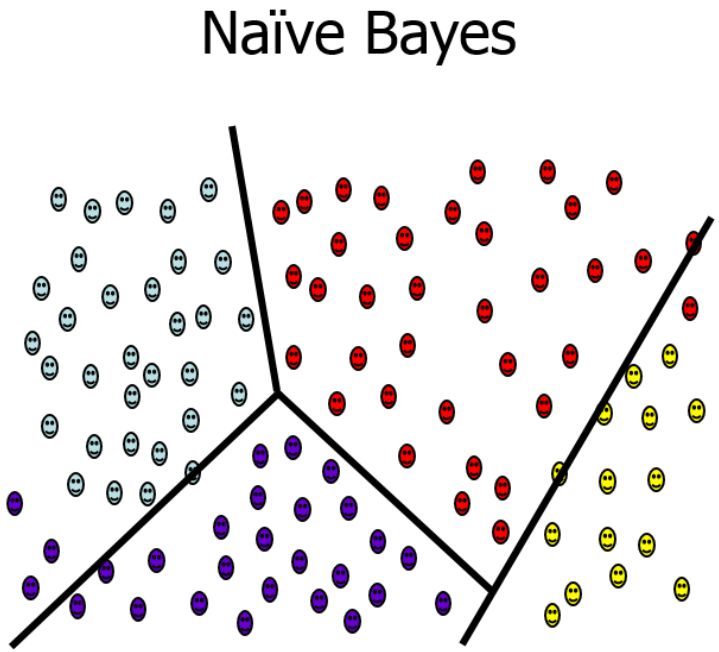
Non Linear (ex: k Nearest Neighbor)



Naïve Bayes

More powerful that Decision Trees





Testing & Evaluation

- After choosing the parameters of the classifiers (i.e. after training it) we need to test how well it's doing on a test set (not included in the training set)
- How trustworthy the model is
- Evaluation can also be an effective tool for guiding us in making future improvements to the model.

Accuracy

- The simplest metric: accuracy, measures the percentage of inputs in the test set that the classifier correctly labeled.
- For example, a spam classifier that predicts correctly spam 60 times in an test set containing 80 email would have an accuracy of $60/80 = 75\%$.
- Important to take into consideration the frequencies of the individual class labels
- If only $1/100$ is spam, an accuracy of 90% is bad
- If $1/2$ is spam, accuracy of 90% is good
- This is also why we use precision & recall and F-measure
- Important: compare with fair baselines

Evaluating classifiers

Contingency table for the evaluation of a binary classifier

	GREEN is correct	RED is correct
GREEN was assigned	a	b
RED was assigned	c	d

- Accuracy = (a+d)/(a+b+c+d)
- Precision: P_GREEN = a/(a+b), P_ RED = d/(c+d)
- Recall: R_GREEN = a/(a+c), R_ RED = d/(b+d)

Representation of Objects

- Each object to be classified is represented as a pair (x, y):
  - where x is a description of the object
  - where y is a label
- Success or failure of a machine learning classifier often depends on choosing good descriptions of objects
  - the choice of description can also be viewed as a learning problem (feature selections)
  - but good human intuitions are often needed here

Feature Engineering

As for many ML tasks, it is possible to generate useful features by hands

- General statistics: text length, text length variance, etc
- Domain or Linguistic dictionaries
- POS, NER tags
- Topic Models as features
- Ad-hoc features: e.g. number of emojis

And also, embeddings (bow, tfidf, distributed vectors, etc) ...

Improving Text Classification Models

Practical tips:

- Text Cleaning : text cleaning can help to reducue the noise present in text data in the form of stopwords, punctuations marks, suffix variations etc.
- More intuitive domain and NLP features along with embedding vectors
- Hyperparamter Tuning in modelling : Tuning the paramters is an important step, a number of parameters such as tree length, leafs, network paramters etc can be fine tuned to get a best fit model.
- Ensemble Models : Stacking different models and blending their outputs can help to further improve the results.

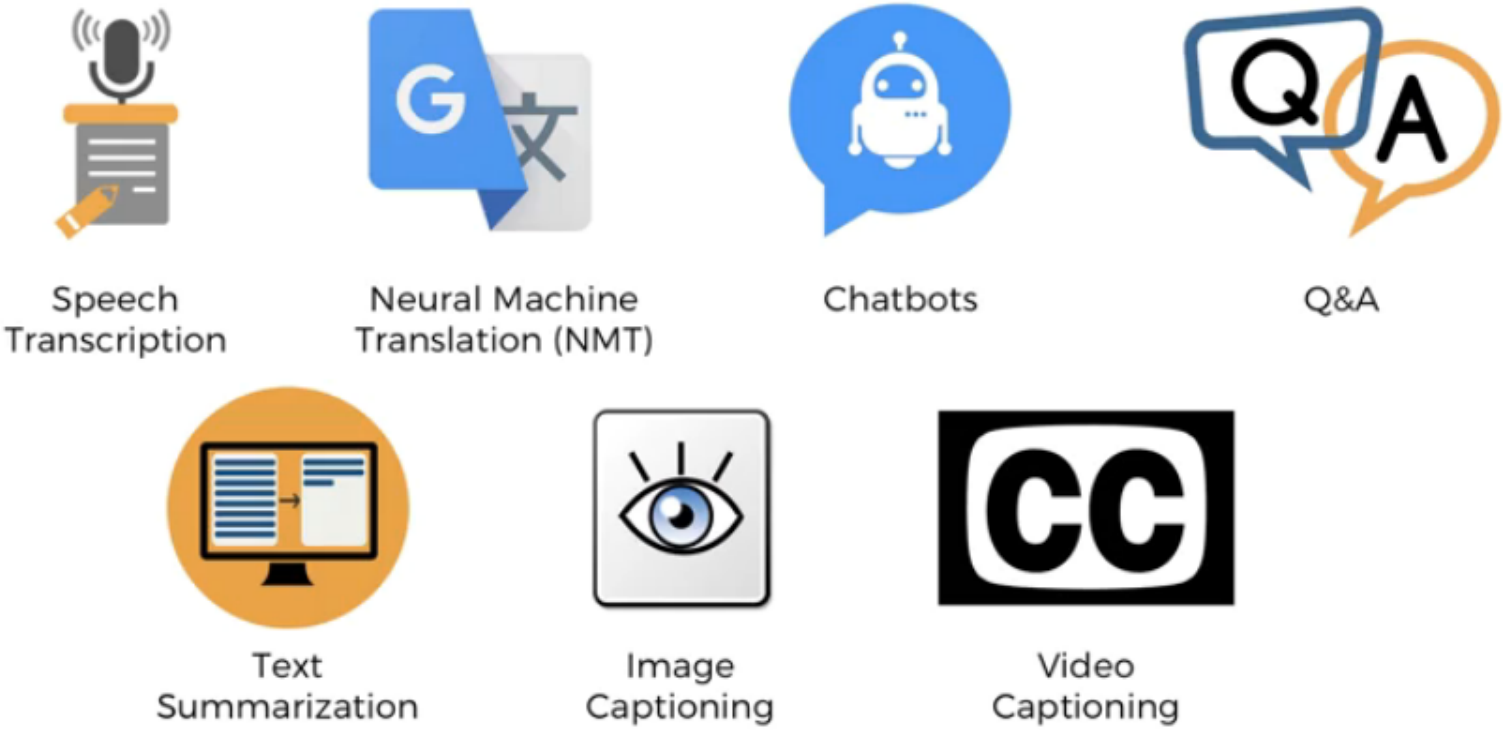
Summary

- We discussed about how to prepare a text dataset
- Perform different types of feature engineering like Count Vector/TF-IDF/ Word Embedding/ Topic Modelling and basic text features,
- Finally trained a variety of classifiers like Naive Bayes/ Logistic regression/ SVM

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