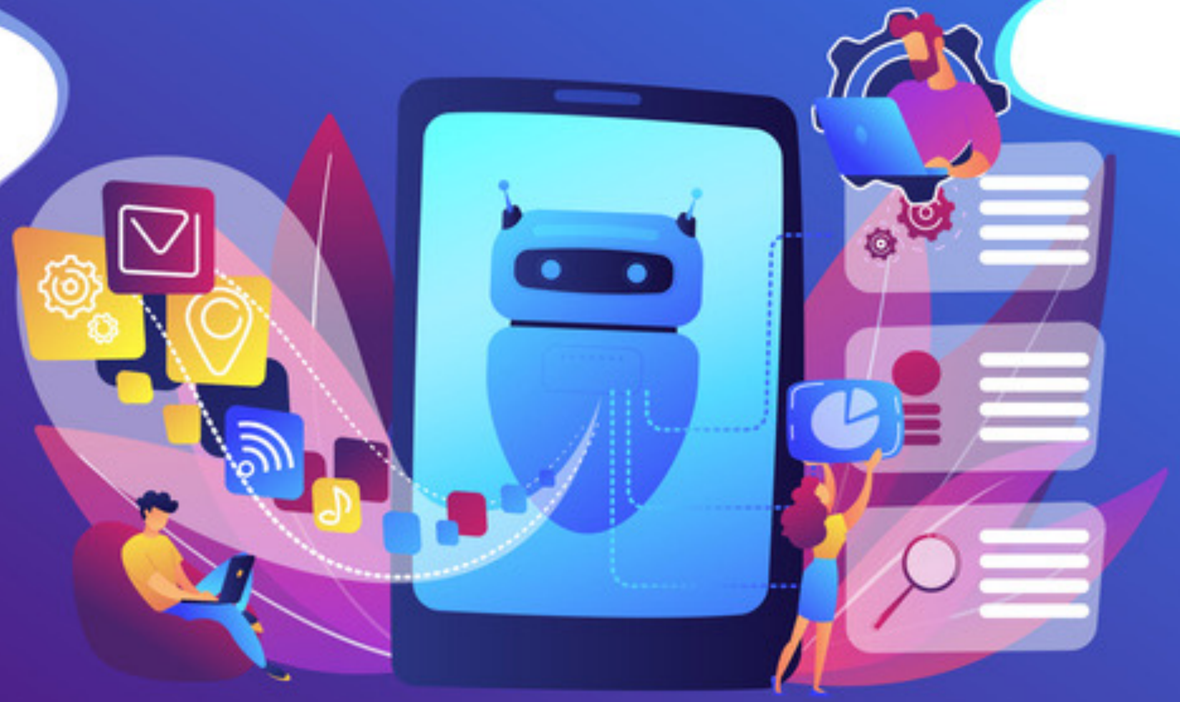


Natural language **processing**

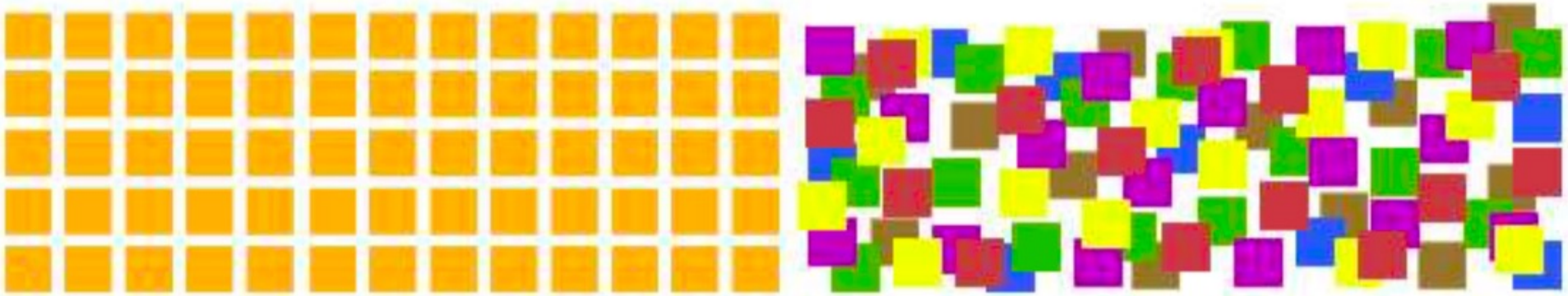
Prof. Dr. Siegfried Handschuh

Zadar 2023

University of St.Gallen, Switzerland

Seth Grimes, a leading industry analyst stated:

"80% of business-relevant information originates in unstructured form, primarily text."



structured data

unstructured data

Overview

1. Motivation & Applications

2. Own Research

3. Hands-On:

   - Word Frequency,

   - Sentiment Analysis,

4. Conclusion

Get to know

- Jupiter Notebook, Python
- notes, data, image
- NLTK, Spacey

Data sets for the exercise

Examples

- Case A: Consumer Complaints (Multiclass)
- Case B: Movie Reviews (Binary Class)
- Case C: Brexit Tweets (Binary Class)

Consumer Complaints

	Unnamed: 0	Product	Consumer complaint narrative
0	549162	Credit card	I recently opened an account with citi simplic...
1	567811	Debt collection	A claim was filed on the original creditor XXX...
2	524772	Debt collection	Disregard the above paragraph, none of your ch...
3	590768	Credit reporting	FRAUDULENT INQUIRES WERE INITIATED ON MY CREDI...
4	221351	Consumer Loan	I was financed by Toyota financial services XX...
5	455688	Debt collection	This company continues to report on my credit ...
6	245659	Money transfers	I found in my email a message from website : X...
7	117564	Credit reporting	Opened an account with XXXX in XX/XX/XXXX -- t...
8	38038	Credit reporting	Transunion charged my credit card without my k...
9	750068	Debt collection	Professional Credit Management of XXXX, AR (P...

Movie Reviews

	Unnamed: 0	rating	text
0	0	pos	in many ways , " twotg " does for tough - guy ...
1	1	neg	it seems like i ' m reviewing cheeseball horro...
2	2	pos	apocalypse now , based on the novel " hearts o...
3	3	pos	known as the most successful , highest - gross...
4	4	neg	" nothing more than a high budget masturbation...
5	5	pos	my summer was recently saved by two very diffe...
6	6	pos	an unhappy italian housewife , a lonely waiter...
7	7	pos	titanic is so close to being the perfect movie...
8	8	neg	the happy bastard ' s quick movie review the c...
9	9	neg	an experience like baby geniuses can have cert...
10	10	neg	well there goes another one . sadly this like ...
11	11	neg	john boorman ' s " zardoz " is a goofy cinemat...
12	12	neg	contrary to popular belief , not every single ...
13	13	neg	it was with great anticipation that i sat down...
14	14	neg	a backdrop of new year ' s eve in 1981 would s...

Fundamental of Natural Language Processing

Motivation

Motivation

Language is multidimensional

Domain knowledge / background knowledge

Rhetoric (also irony) / argumentation

Speech Acts

Sentiment/Opinion

Metaphors/Idioms

cultural contexts

Linguistics (e.g. dialects)

Phonetics

…

[Handschuh et. al: KCap 2001]

[Handschuh et. al: EKAW 2002]

[Cimiano, Handschuh, Staab: WWW 2004]

[Handschuh: Thesis 2005]

[Cimiano, Handschuh: 2003]

[Groza, Handschuh, Decker: 2010]

[Groza, Handschuh, Bordea: ACM 2010]

How hard is it?

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Some common NLP tasks

- **Classifying whole sentences**: Getting the sentiment of a review, detecting if an email is spam, determining if a sentence is grammatically correct or whether two sentences are logically related or not

- **Classifying each word in a sentence**: Identifying the grammatical components of a sentence (noun, verb, adjective), or the named entities (person, location, organisation)

- **Generating text content**: Completing a prompt with auto-generated text, filling in the blanks in a text with masked words

- **Extracting an answer from a text**: Given a question and a context, extracting the answer to the question based on the information provided in the context.

- **Generating a new sentence from an input text**: Translating a text into another language, summarising a text

Information Extraction & Sentiment Analysis

Attributes:
- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

✓ • nice and compact to carry!

✓ • since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!

✗ • the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafsky

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.
-Chris

Event: **Curriculum mt**
Date: **Jan-16-2012**
Start: **10:00am**
End: **11:30am**
Where: **Gates 159**

Ambiguity makes Text Mining hard: “Crash blossoms”

Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Why else is text mining difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

But that's what makes it fun!

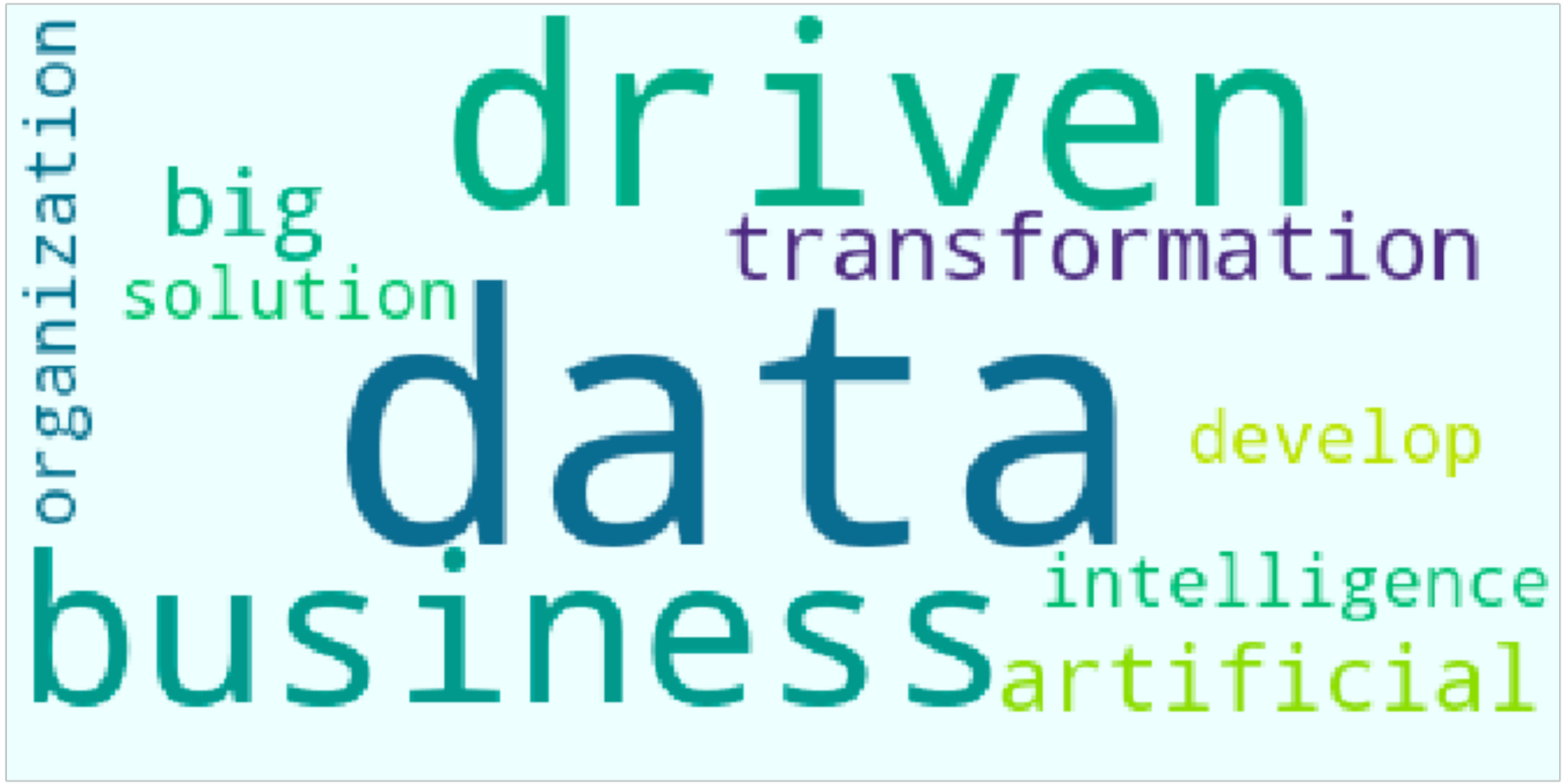
Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources

- How we generally do this:
  - probabilistic models built from language data
    - P(“maison” → “house”) high
    - P(“L’avocat général” → “the general avocado”) low
  - Luckily, rough text features can often do half the job.

Hands On

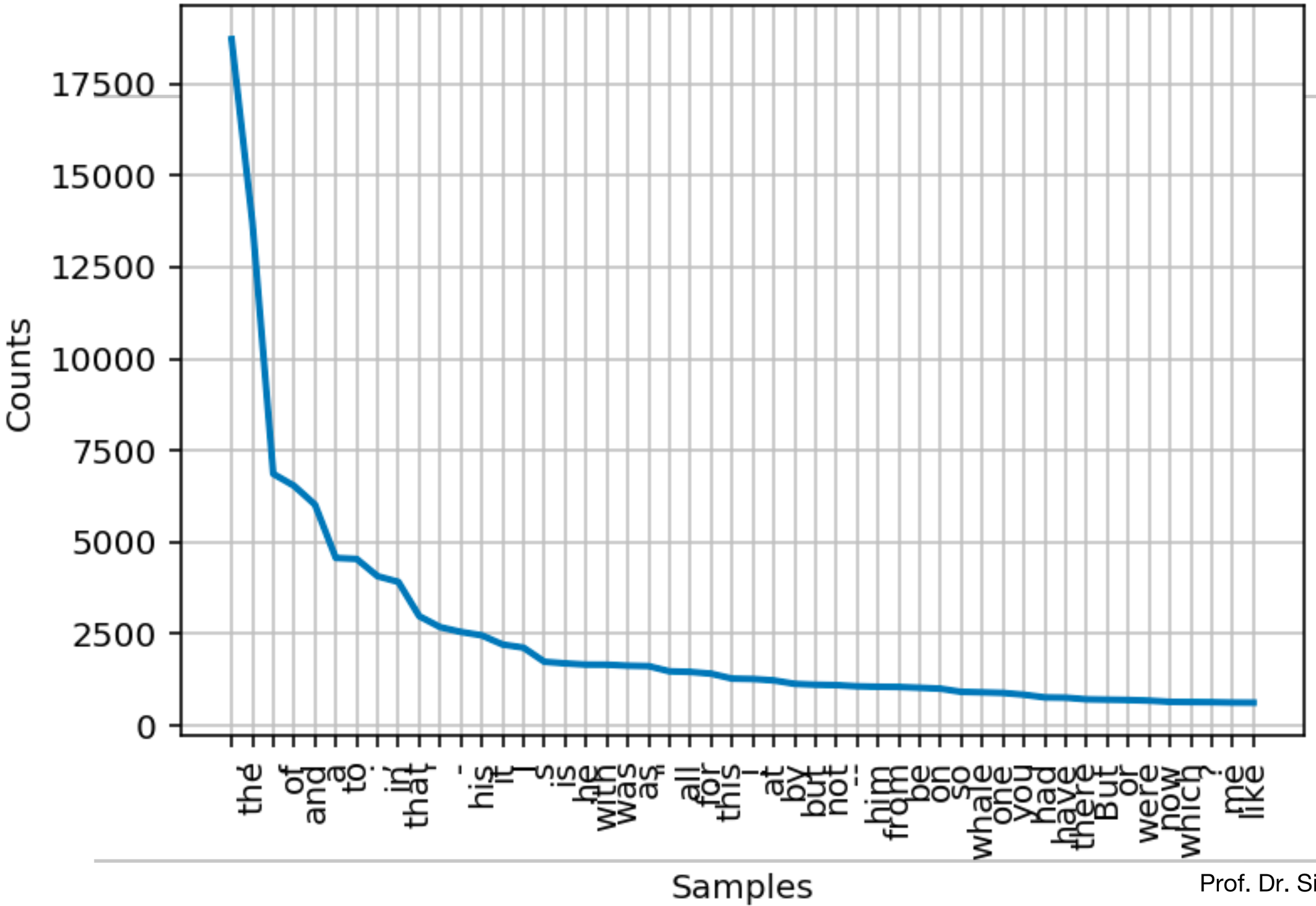
Word Frequency



Frequency Distributions

Word Tally

the	𝍸 𝍸 𝍸 ||||
been	𝍸 𝍸 |
message	||||
persevere	|
nation	𝍸 |||



Word frequency within “Moby Dick”

Definitions

- Word – A delimited string of characters as it appears in the text.
- Term – A "normalized" word (case, morphology, spelling etc); an equivalence class of words.
- Token – An instance of a word or term occurring in a document.
- Type – The same as a term in most cases: an equivalence class of tokens.

Stopwords

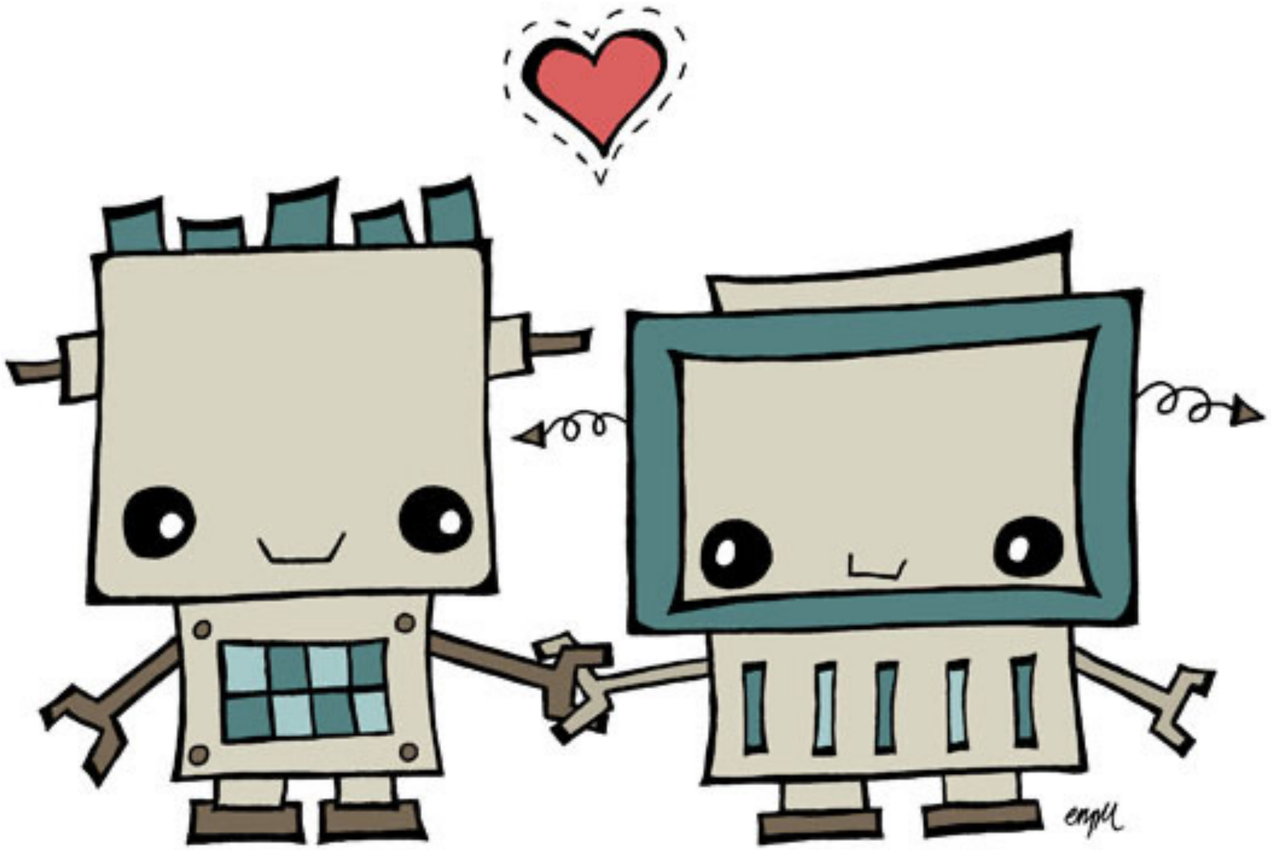
- Stop words are words which are filtered out before processing of natural language data.
- Stop words are generally the most common words in a language; there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list.
- Some tools avoid removing stop words to support phrase search.

Word Cloud

- A text cloud or word cloud is a visualization of word frequency in a given text as a weighted list.
- Word clouds have been subject of investigation in several usability studies.
  - Large tags attract more user attention than small tags (effect influenced by further properties, e.g., number of characters, position, neighboring tags).
  - Scanning: Users scan rather than read tag clouds.
  - Centering: Tags in the middle of the cloud attract more user attention than tags near the borders (effect influenced by layout).

Sentiment Analysis

Goal: Giving emotion understanding to machines



Positive or negative movie review?

- unbelievably disappointing

- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

- It was pathetic. The worst part about it was the boxing scenes.

Google Product Search

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, **$100** nearby ★★★★☆ 377 reviews
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

1 star	2	3	4 stars	5 stars

What people are saying

ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints came out with great quality."

Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

Product summary | Find best price | **Customer reviews** | Specifications | Related items

$121.53 - $242.39 (14 stores)

☐ Compare

Average rating ★★★½☆ (144)

★★★★★	(55)
★★★★☆	(54)
★★★☆☆	(10)
★★☆☆☆	(6)
★☆☆☆☆	(23)
☆☆☆☆☆	(0)

Most mentioned

Performance	(57)
Ease of Use	(43)
Print Speed	(39)
Connectivity	(31)

More ▼

Show reviews by source

- Best Buy (140)
- CNET (5)
- Amazon.com (3)

Sentiment analysis has many other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

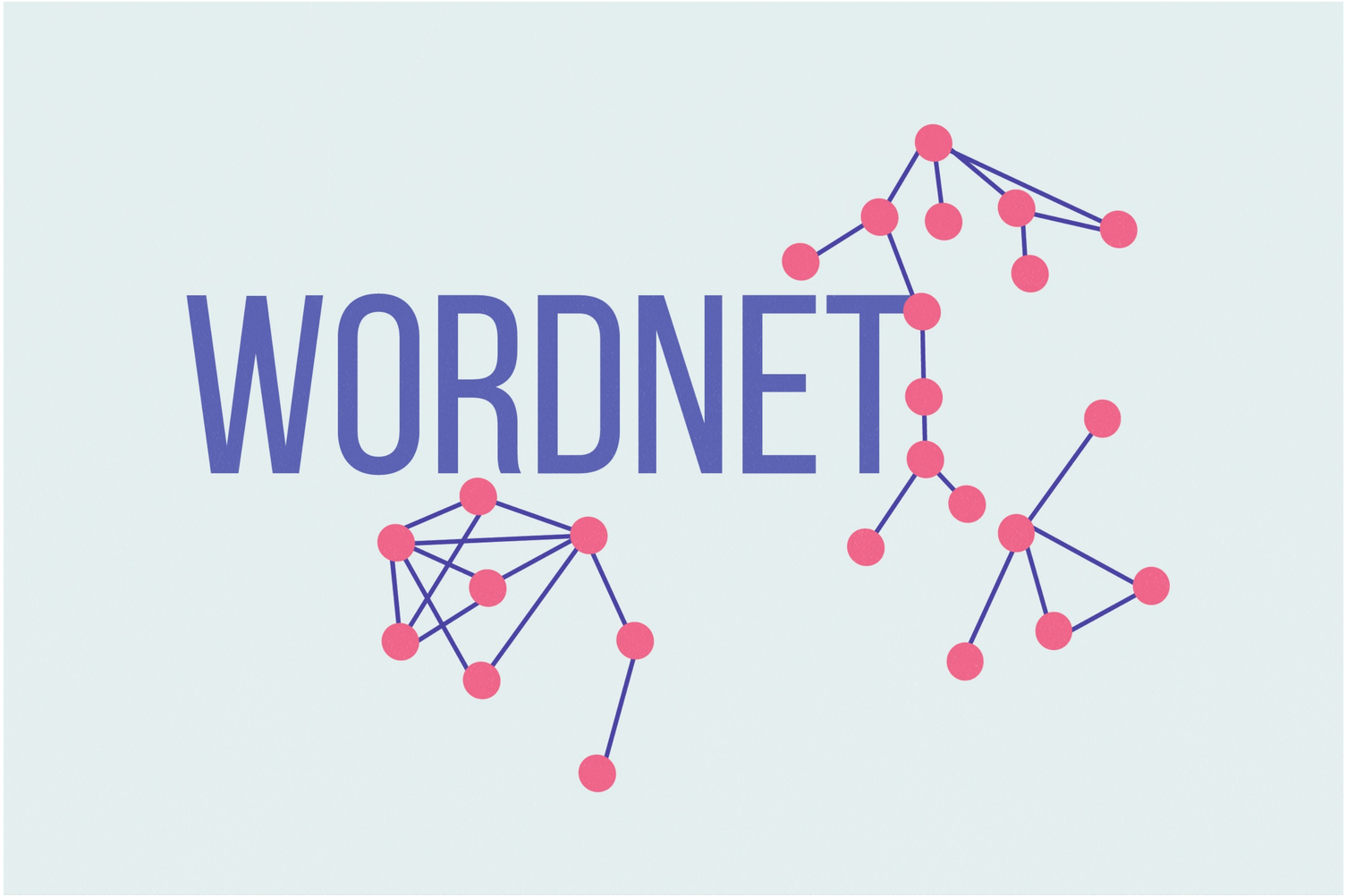
Why sentiment analysis?

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence? Is despair increasing?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

Sentiment Analysis

- Simplest task:
  - Is the attitude of this text positive or negative?
- More complex:
  - Rank the attitude of this text from 1 to 5
- Advanced:
  - Detect the target, source, or complex attitude types

SentiWordNet



Part of Speech

- Category of words which have similar grammatical properties.
- Words that are assigned to the same word part of speech generally display similar behaviour in terms of syntax.
- Also referred to as: lexical categories, word classes, morphological classes, lexical tags.

Dionysius Trax of Alexandria [c. 100 BC] describes 8 parts-of-speech:

- Noun
- Verb
- Pronoun
- Preposition
- Adverb
- Conjunction
- Participle
- Article

Examples of POS Tags

Noun	book/books, nature, Germany
Verb	eat, wrote
Auxiliary	can, should, have
Adjective	new, newer, newest
Adverb	well, urgently
Number	872, two, first
Article/Determiner	the, some
Conjunction	and, or
Pronoun	he, my
Preposition	to, in
Particle	off, up
Interjection	Yay, Yeah, Zing

Opinion related Properties

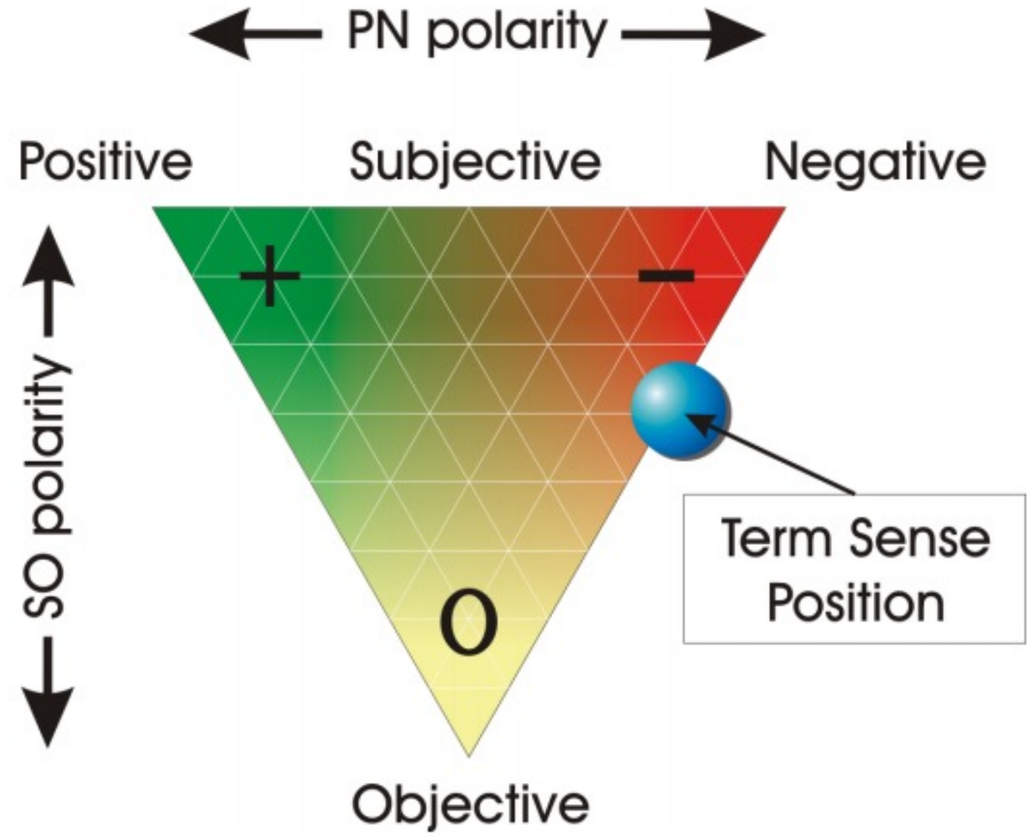


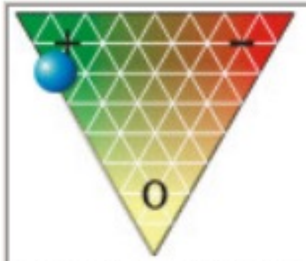
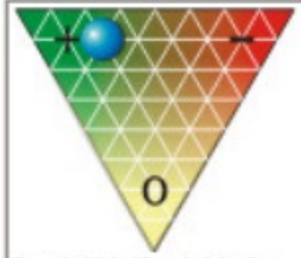
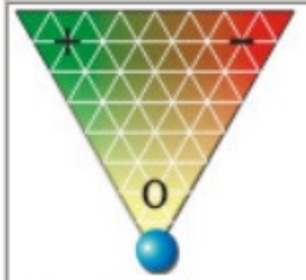
Figure 1. The graphical representation adopted by SENTIWORDNET for representing the opinion-related properties of a synset.

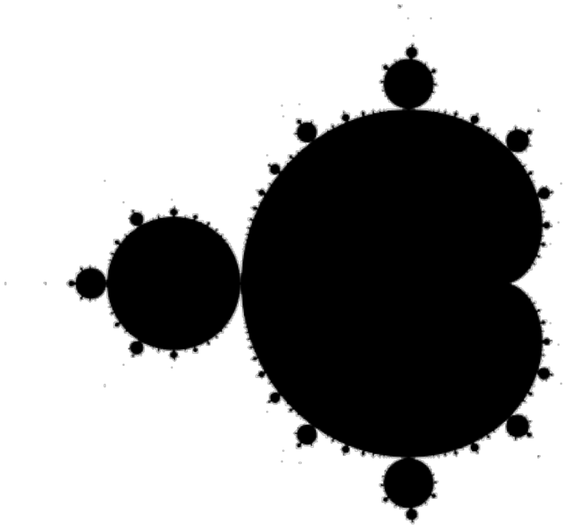
Example

estimable | Search word | show position

Adjective

3 senses found.

 P = 0.75, N = 0, O = 0.25	estimable(1) *deserving of respect or high regard*
 P = 0.625, N = 0.25, O = 0.125	honorable(5) good(4) respectable(2) estimable(2) *deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"*
 P = 0, N = 0, O = 1	computable(1) estimable(3) *may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"*



TextBlob

TextBlob Sentiment Analysis

- TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

- The analysis returns two values: polarity and subjectivity. From what I read online, the polarity score is a float within the range [-1.0, 1.0] where 0 indicates neutral, +1 a very positive attitude and -1 a very negative attitude. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

It’s not always as easy as it looks

“Rubbish hotel in Madrid”

What makes reviews hard to classify?

- Subtlety:
  - Perfume review in *Perfumes: the Guide*:
    - “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
  - Dorothy Parker on Katherine Hepburn
    - “She runs the gamut of emotions from A to B”

Thwarted Expectations and Ordering Effects

- “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it **can’t hold up**.”
- Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is **not so good** either, I was surprised.

Conclusion

- Natural Language Toolkit for Beginners
  - Tokenisation, stop word removal, lemmatisation, stemming, part of speech tag
- Word Frequency
  - data preparation, Long tail distribution of words in a document
- Wordcloud
- Sentiment Analysis

[('the', 'DT'), ('dogs', 'NNS'), ('are', 'VBP'), ('barking', 'VBG'), ('outside', 'IN'), ('.', '.')]