WORKING WITH TEXT

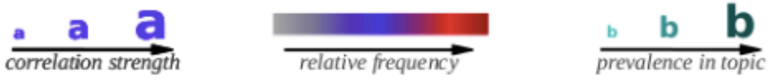
BEYOND WORD CLOUDS

PRESENTED BY ANNA NICANOROVA


bigdata
training
showcase
tutorials
thursdays
microsoft
day
enterprise
thecube
enter
time
come
hortonworks
keynotes
tomorrow
new
informatica
database
amitwalia
including
free
json
infabd
join
get
first
miss
using
jimmchugh
can
stop
chat
nyc
build
boost
make
world
ready
find
need
schedule
w
help
keynote
watch
dj44
livedemo
azure
1m
10115
become
going
learn
social
look
big
algorithms
datakind
today
via
kdnuggets
337
merv
insight
check
win
meetlike
session
apachespark
streaming
building
learning
take
machine
bigdatanyc
great
million
one
cloudera
wish
spark
now
datasciencejust
people
interested
bigdataready
visit
jakeporway
good
analytics
makes
insights
mapr
thoughtspot
talk
mongodb
data
us
hadoop
things
booth
available
challenge


bestie
she's
can't_believe
blessed
daughter
so_tired
her_first
soooo
she
cry
mom
loving
beautiful
princess
cute
yummy
hubby
sooooo
super_excited
time_with
her
loved
cake
wishes_she
:d_<3
so_much
thank_you
lots
husband
loves_her
girl
love_you
so_happy
soo
such_a
nails
dress
I_miss
sick
thankful
ugh
my_heart
christmas
am_so
wonderful
:(
excited
I'm_so
drama
my_little
love_him
to_see_my
gosh
love_you_guys
having_a
proud_of
yum
tummy
?_?
mommy
aunt
much_fun
my_hair
:)
shopping
birthday
smile
bed
herself
boyfriend
sad
<3
yay_!
like_it_on
:)
girls
can't_wait
lovely
her_new
make_me
so_glad
sooo
dinner
best_friend
sister
cookies
babysitting
omg
my_family
chocolate
love_them
baking
baby
laugh
with_my
daddy
go_away
adorable
babies
cleaning
Penn | World Well-Being Project | wwbp.org


windows_7
2-0
dick
feature
nfl
holy_shit
www
bastard
lebron
weapons
gun
political
modern_warfare
beard
fans
his_own
fifa
his_mind
gaming
victory
playoffs
government
chuck_norris
win
fucked
fuck
pc
my_wife
master
http://
religion
bass
fuck_yeah
espn
software
champions
hell
hit_me_up
black_ops
vs
world_cup
league
system
team
beat
my_girlfriend
youtube.com
shave
epic
his_phone
india
nation
destroyed
1-0
call_of_duty
fuckin
speed
war
cod
fucking
wishes
nba
he
ftw
god_damn
beer
nahi
wins
politics
shaved
thinks_he
xbox
championship
himself
360
nuclear
engineering
of_death
metal
ps3
teams
his_life
cock
album
halo
creed
sword
match
legendary
arsenal
shit
players
loves_his
brazil
legend
united
president
player
germany
Penn | World Well-Being Project | wwbp.org


a
a
a
correlation strength
relative frequency
b
b
b
prevalence in topic

SpaceAngel #StopBrexit & #GTTO @spaceangel1964 · May 11

Please contribute

I want to create a **word cloud** which captures the words/phrases that sum up why you believe we should Remain in the EU.

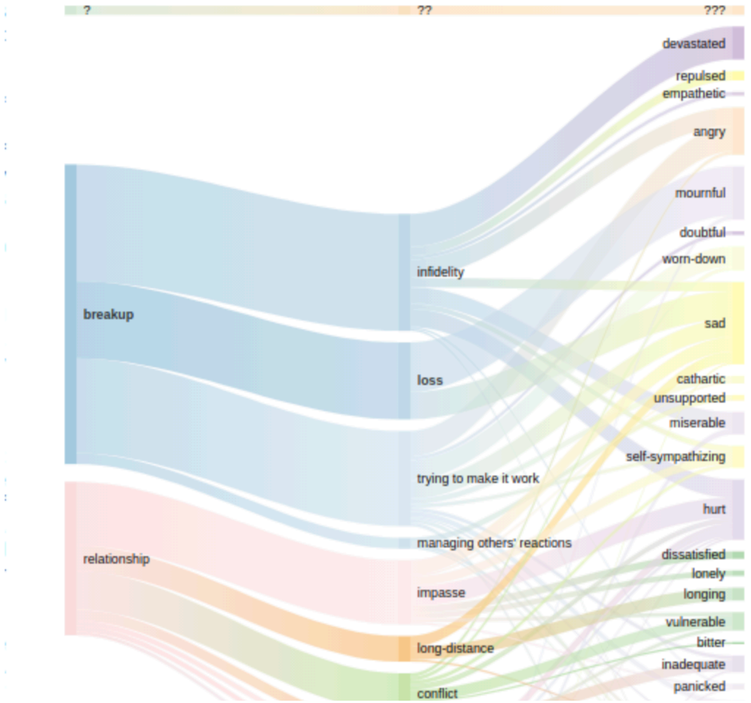
We can use it in the run up to #EuroElections2019 - ideally to reach outside our bubble! Please add your words below by Reply & RT 🌌😇🇪🇺

860 369 439

qxzito @qxzito1 · 9h

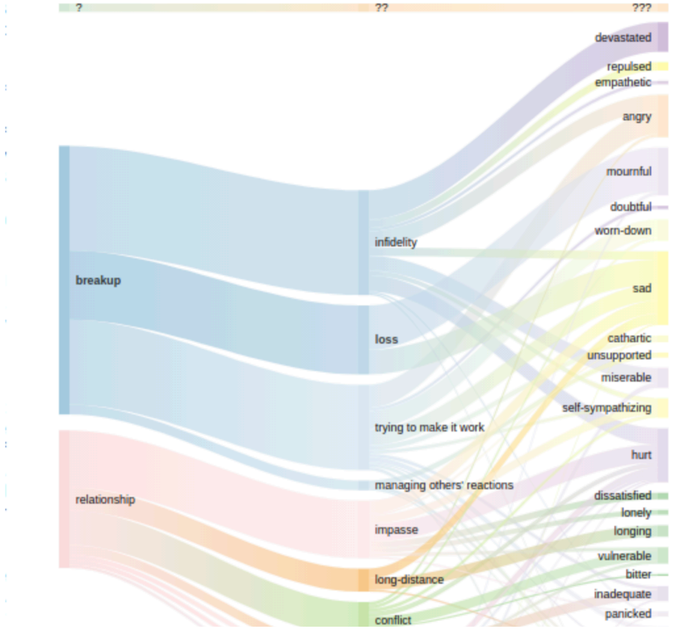
Tue 05.14.19 NY Times front page as a **Word Cloud**

ALTERNATIVES TO WORD CLOUDS

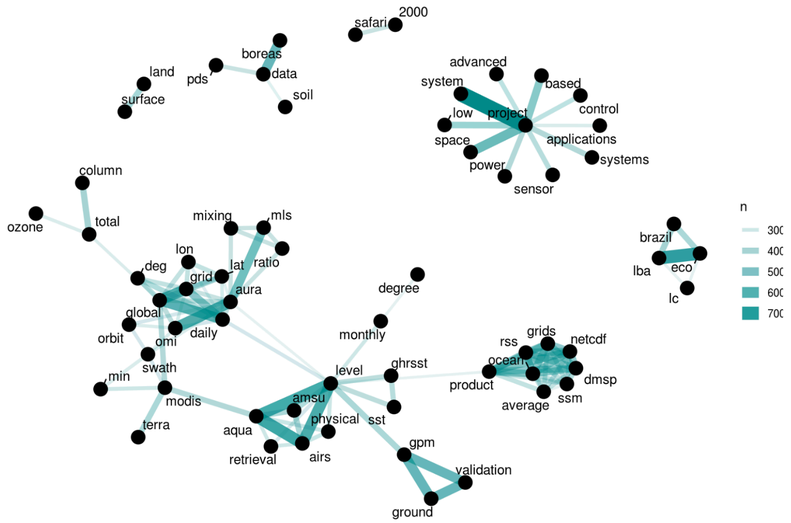


SANKEY DIAGRAM

ALTERNATIVES TO WORD CLOUDS



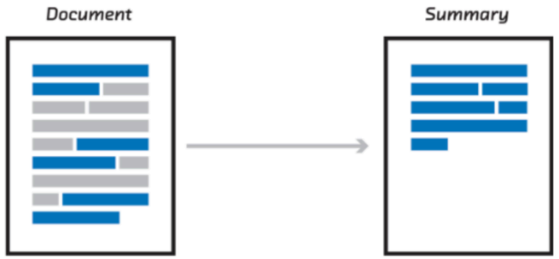
SANKEY DIAGRAM



WORD NETWORKS

See also: http://textexture.com/

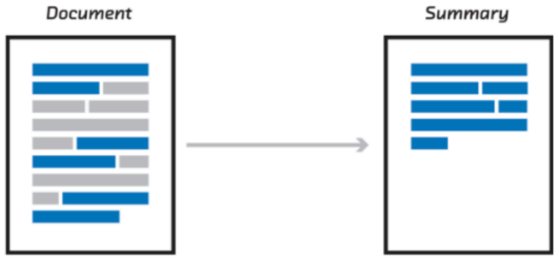
ALTERNATIVES TO WORD CLOUDS



TEXT SUMMARIZATION

https://towardsdatascience.com/can-we-please-stop-using-word-clouds-eca2bbda7b9d

ALTERNATIVES TO WORD CLOUDS



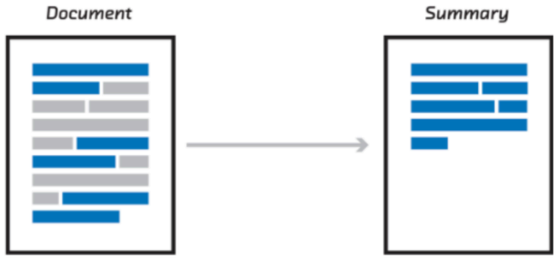
TEXT SUMMARIZATION

INFO HIGHLIGHT

(SPARKLINE/CALLOUT)

https://towardsdatascience.com/can-we-please-stop-using-word-clouds-eca2bbda7b9d

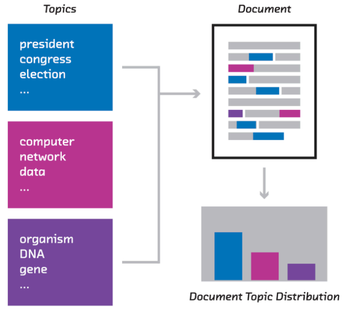
ALTERNATIVES TO WORD CLOUDS



TEXT SUMMARIZATION

INFO HIGHLIGHT

(SPARKLINE/CALLOUT)



TOPIC MODELING

https://towardsdatascience.com/can-we-please-stop-using-word-clouds-eca2bbda7b9d

ELEPHANT IN THE ROOM

ELEPHANT IN THE ROOM

AI

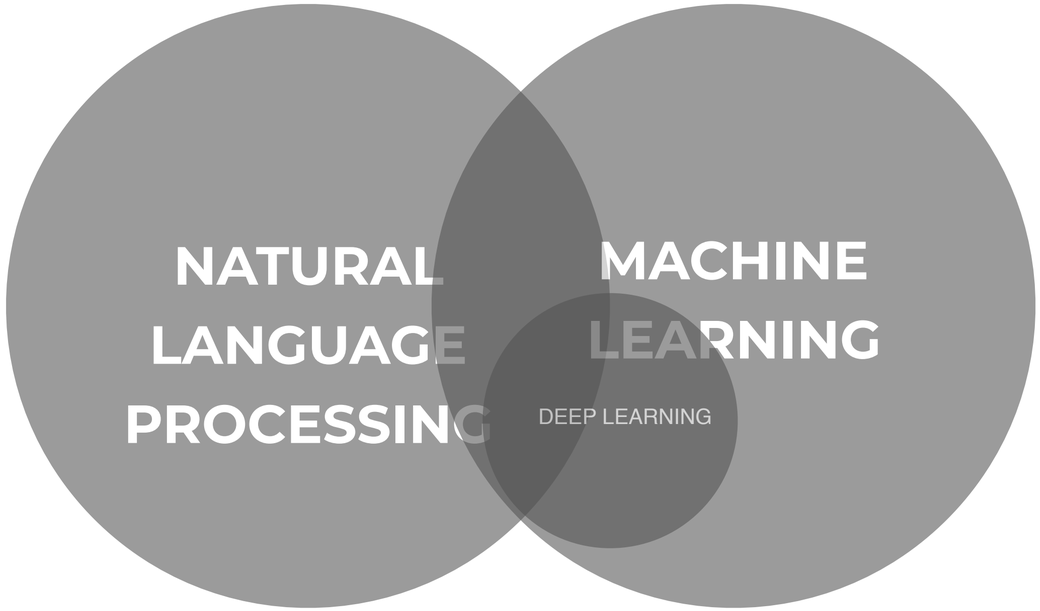
TEXT IS HARD

Consider robots

- They can walk (now run, and maybe even climb stairs)
- They can play chess
- They can play football
- They can recognize people by face
- ⊗ They can't hold a conversation
  - Language turns out to be very difficult

SOURCE: http://compling.hss.ntu.edu.sg/courses/hg8003.2014/pdf/wk-01.pdf

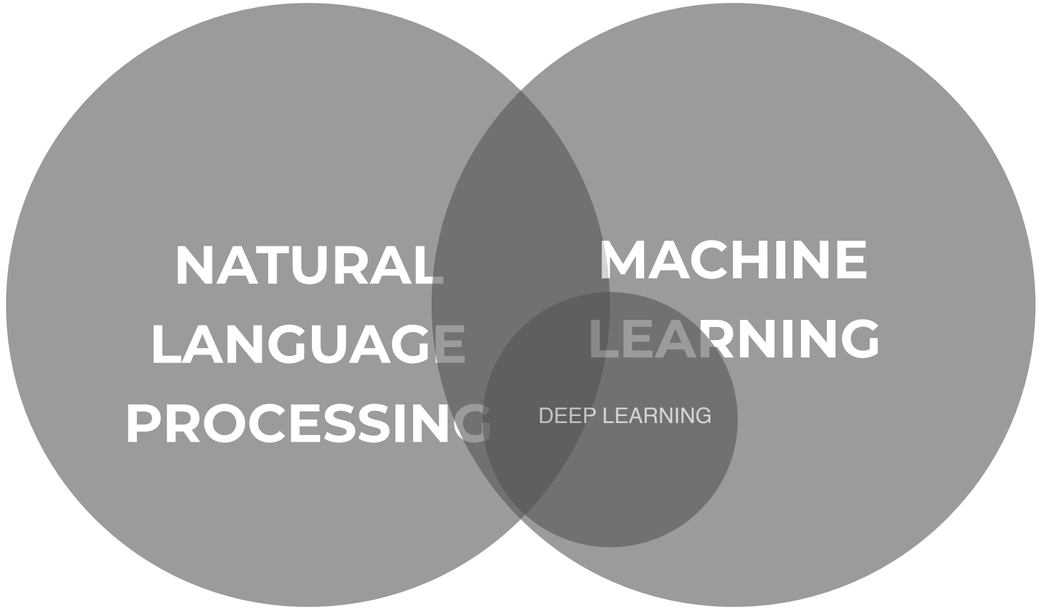
ML/NLP/AI...



SOURCE: Chapter 9 : Natural Language Processing

ML/NLP/AI...

- Syntax – What part of given text is grammatically true.
- Semantics – What is the meaning of given text?
- Pragmatics – What is the purpose of the text?
- Phonology – It is systematic organization of sounds in language.
- Morphology – It is a study of words formation and their relationship with each other.



- Text classification and clustering
- Information retrieval and extraction
- Machine translation(one language to another)
- Question and answering system
- spelling and grammar checking
- Topic modeling and sentiment analysis
- Speech recognition

SOURCE: Chapter 9 : Natural Language Processing

LANGUAGE TECHNOLOGY

MOPSTLY SOLVED

SPAM DETECTION

- Let’s go to Agra!
- Buy VIAGRA

PART OF SPEECH TAGGING

- Colorless (ADJ) green (ADJ) ideas (NOUN) sleep (VERB) furiously (ADV)

NAMED ENTITY RECOGNITION

- Einstein (Person) met with UN(ORG) officials in Princeton (Location)

** as of April 2019

SOURCE: quora

LANGUAGE TECHNOLOGY

MOPSTLY SOLVED

SPAM DETECTION

- Let's go to Agra!
- Buy VIAGRA

PART OF SPEECH TAGGING

- Colorless (ADJ) green (ADJ) ideas (NOUN) sleep (VERB) furiously (ADV)

NAMED ENTITY RECOGNITION

- Einstein (Person) met with UN(ORG) officials in Princeton (Location)

MAKING GOOD PROGRESS

SENTIMENT ANALYSIS

- Best pizza in New York
- We waited for 30 minutes to get seated

MACHINE TRANSLATION

- I like learning chinese
- 我喜欢学中文
- Wǒ xǐhuān xué zhōngwén

INFORMATION EXTRACTION

- You're all booked for your flight NYC to SF on May 6th
- Fly May 6th. Add to calendar?

WORD SENSE DISAMBIGUATION

- I need batteries for my *mouse*

** as of April 2019

SOURCE: quora

LANGUAGE TECHNOLOGY

MOPSTLY SOLVED

SPAM DETECTION

- Let's go to Agra!
- Buy VIAGRA

PART OF SPEECH TAGGING

- Colorless (ADJ) green (ADJ) ideas (NOUN) sleep (VERB) furiously (ADV)

NAMED ENTITY RECOGNITION

- Einstein (Person) met with UN(ORG) officials in Princeton (Location)

MAKING GOOD PROGRESS

SENTIMENT ANALYSIS

- Best pizza in New York
- We waited for 30 minutes to get seated

MACHINE TRANSLATION

- I like learning chinese
- 我喜欢学中文
- Wǒ xǐhuān xué zhōngwén

INFORMATION EXTRACTION

- You're all booked for your flight NYC to SF on May 6th
- Fly May 6th. Add to calendar?

WORD SENSE DISAMBIGUATION

- I need batteries for my *mouse*

STILL REALLY HARD

QUESTION/ANSWER

- How effective is Ibuprofen in reducing fever in patients with acute febrile illness?

PARAPHRASING

- 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

DIALOGUE

- Where is Citizen Kane is playing in NYC?
- Angelica theather at 7:30. Do you want a ticket?

** as of April 2019

SOURCE: quora

PAID NLP API'S

You don't really need to write your own code - there are plenty providers, who for a certain charge will process text for you

MORE HERE: https://blog.rapidapi.com/best-nlp-api/#10-duckduckgo-zero-click-info-api

NLP IN PYTHON

Python is by far the most popular language for doing Natural Language Processing. This list is constantly updated as new libraries come into existence.

Natural Language ToolKit

python

gensim
topic modelling for humans

spaCy

scikit learn

TextBlob

Stanford University
S
NLP
Natural Language Processing

PyText NLP

Sumy python

flair

Good overview & comparison here: https://nlpforhackers.io/libraries/

TUTORIAL PLAN

STEP 1:

CLEANING DATA

STEP 2:

TEXT EXPLORATION

STEP 3:

ENTITY EXTRACTION

STEP 4:

SUMMARIZATION

STEP 5:

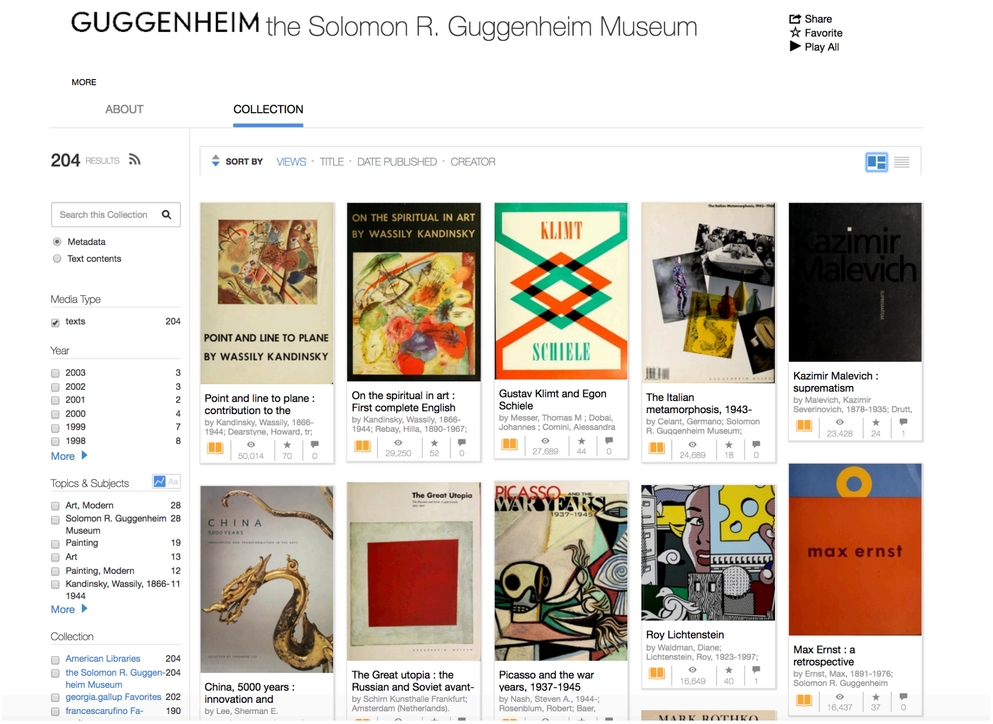
TOPING MODELING

STEP 6:

GETTING FANCY &

CREATIVE APPROACH TO SHOWING DATA RESULTS

ART BOOKS FROM GUGGENHEIM MUSEUM



A NOTE ON SCRAPING

```
def get_book_list():
    '''get book list available from archives
    # there is infinite scroll on the page, so use this to load more: https://www.accordbox.com/blog/how-crawl-infinite-scrolling-pages-using-python/
     '''
    global alldata
    npages = 3
    alldata = pd.DataFrame(columns=[])

    for pagenum in range(1,npages):
        url = "https://archive.org/details/guggenheimmuseum?and%5B%5D=mediatype%3A%22texts%22&sort=titleSorter&page=" + str(pagenum)
        page = requests.get(url)
        soup = BeautifulSoup(page.text, 'html.parser')
        books = soup.find_all("div", {"class": "item-ia"})
        data = pd.DataFrame(columns=[])
        data['html'] = books
        alldata = alldata.append(data)
```

STEP 0
SET-UP

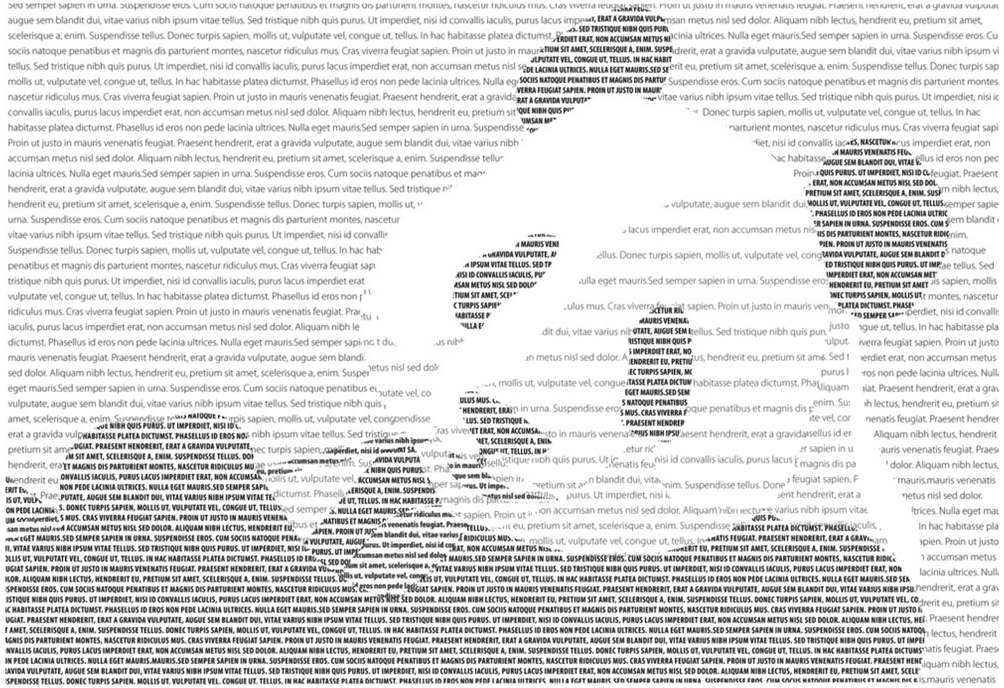
EXECUTABLE NOTEBOOK (IF YOU WANT TO RUN REMOTE):

https://notebooks.azure.com/anna-nicanorova/projects/wcaiconf-2019

GITHUB CODE (IF YOU WANT TO RUN LOCALLY): https://github.com/AnnaNican/wcaiconf_2019

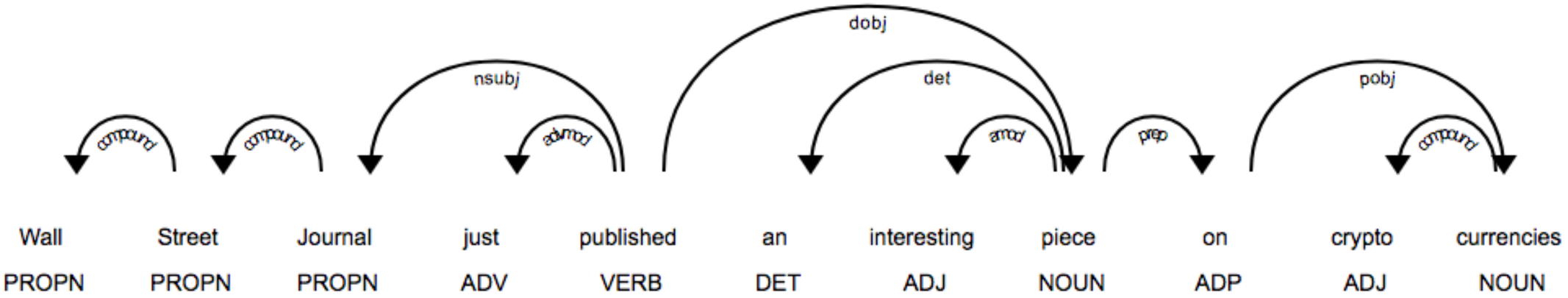
TEXT FILES : https://github.com/AnnaNican/wcaiconf_2019/tree/master/data/books

STEP 1
LOAD AND CLEAN DATA

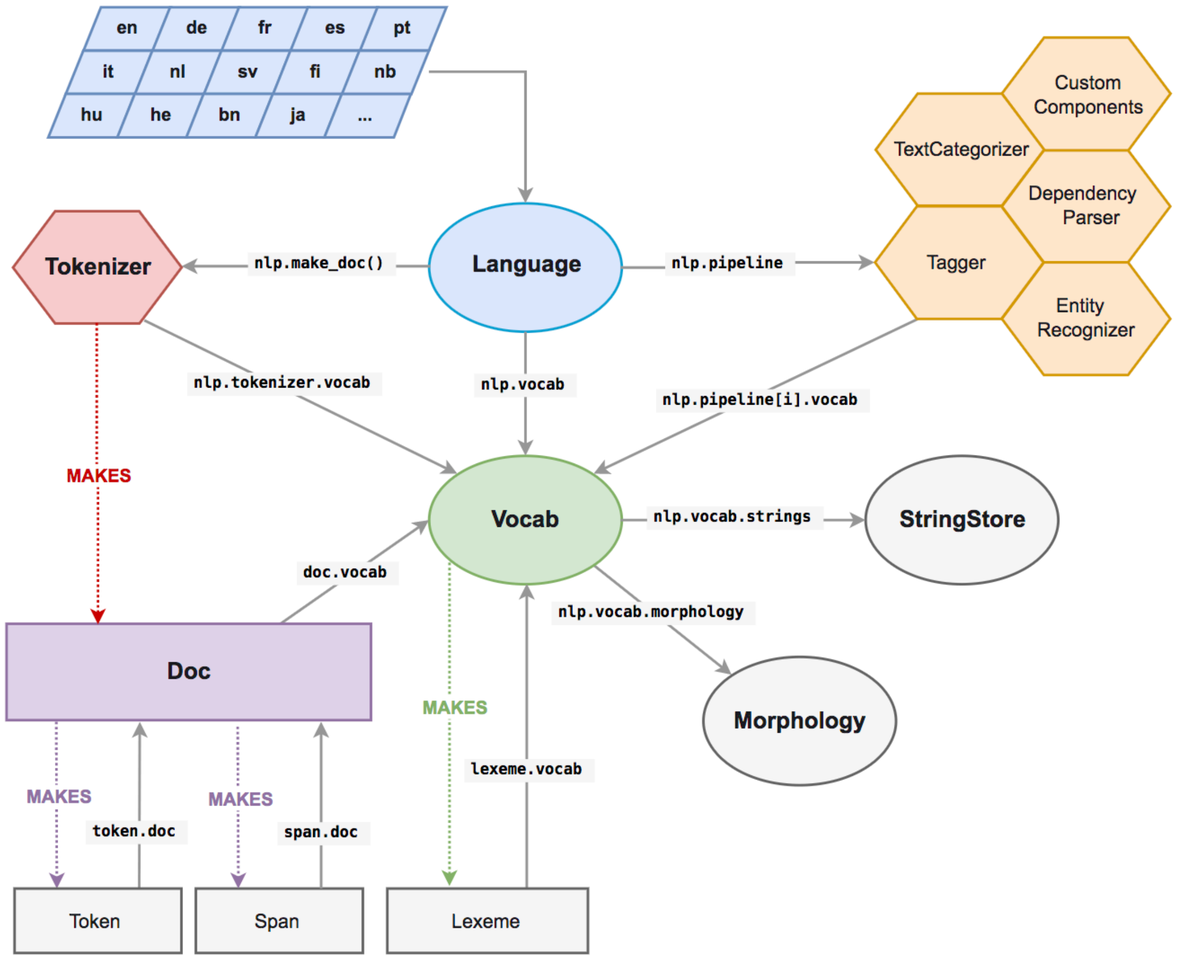


	ASCII	ISO-8859-15 (latin-9)	CP-1252 (Windows 1252)	UTF-8
a	01100001	01100001	01100001	01100001
€	NA	10100100	10000000	11100010 10000010 10101100
¤	NA	NA	10100100	11000010 10100100

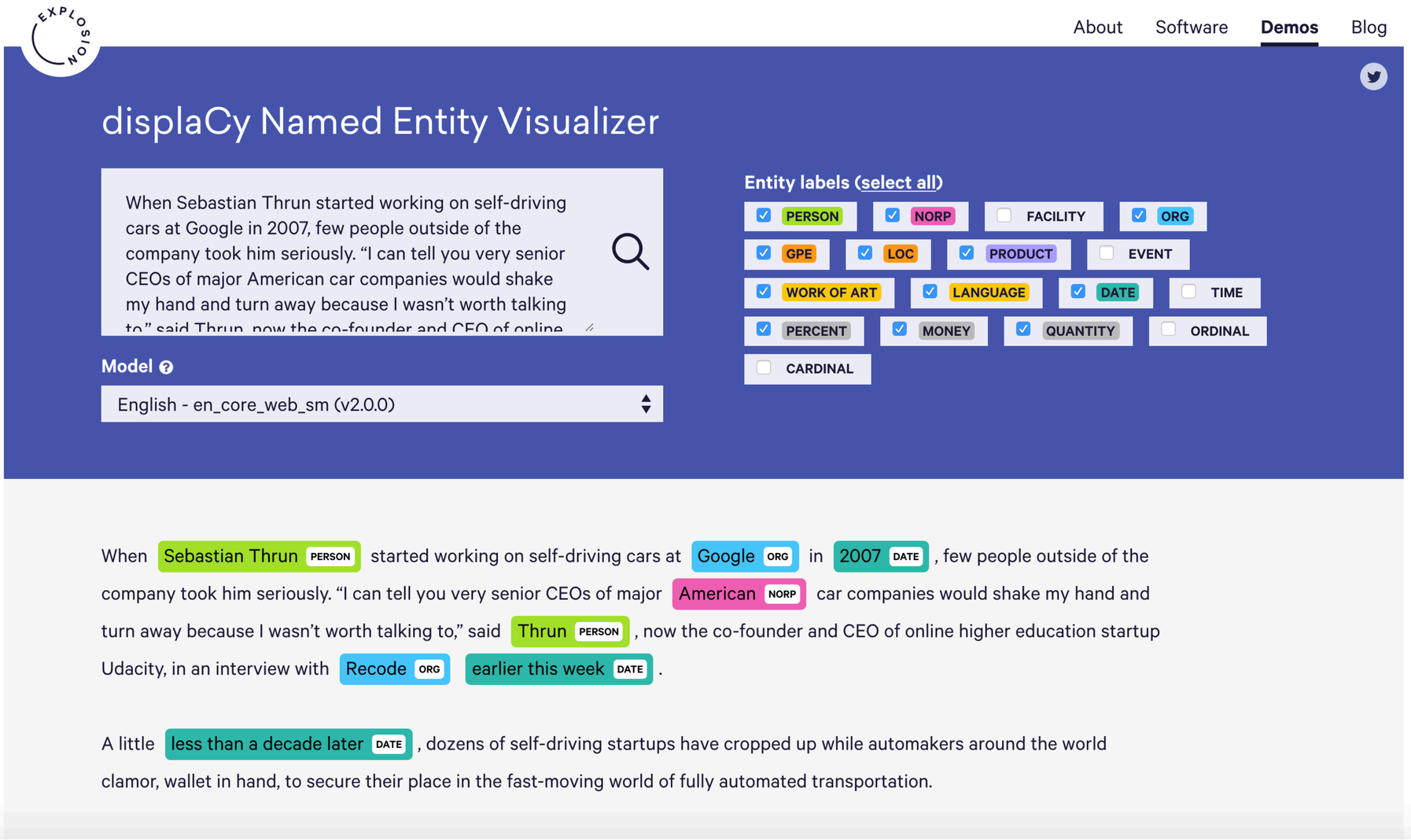
STEP 2
TEXT EXPLORATION



STEP 2
TEXT EXPLORATION



STEP 3
ENTITY EXTRACTION



STEP 4
TEXT SUMMARIZATION

Extractive Summarization

Select parts (typically sentences) of the original text to form a summary.

- Easier
- Too restrictive (no paraphrasing)
- Most past work is extractive

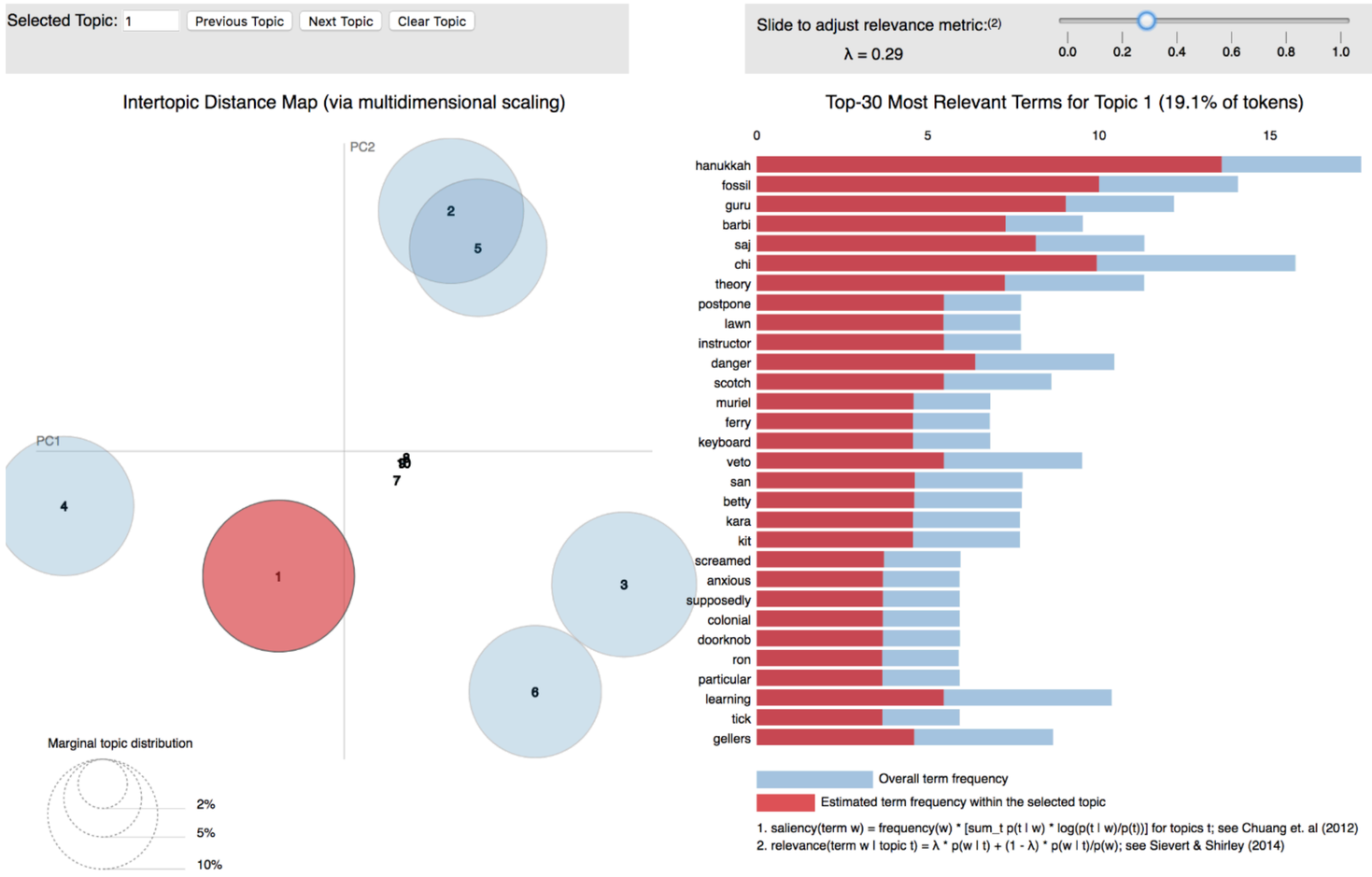
Abstractive Summarization

Generate novel sentences using natural language generation techniques.

- More difficult
- More flexible and human
- Necessary for future progress

Source: https://www.slideshare.net/aclanthology/abigail-see-2017-get-to-the-point-summarization-with-pointergenerator-networks

STEP 5
TOPIC MODELING

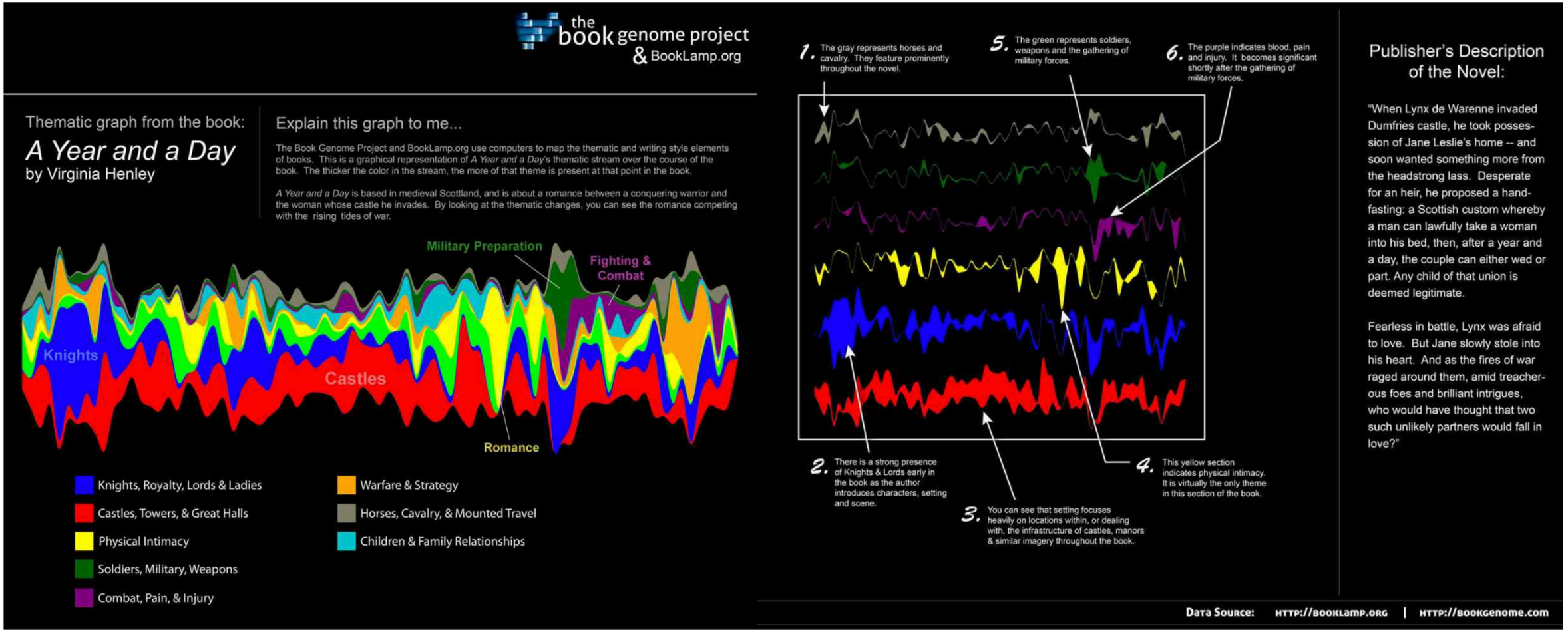


Source: https://medium.com/@sherryqixuan/topic-modeling-and-pyldavis-visualization-86a543e21f58

STEP 6
BEYOND WORD CLOUDS

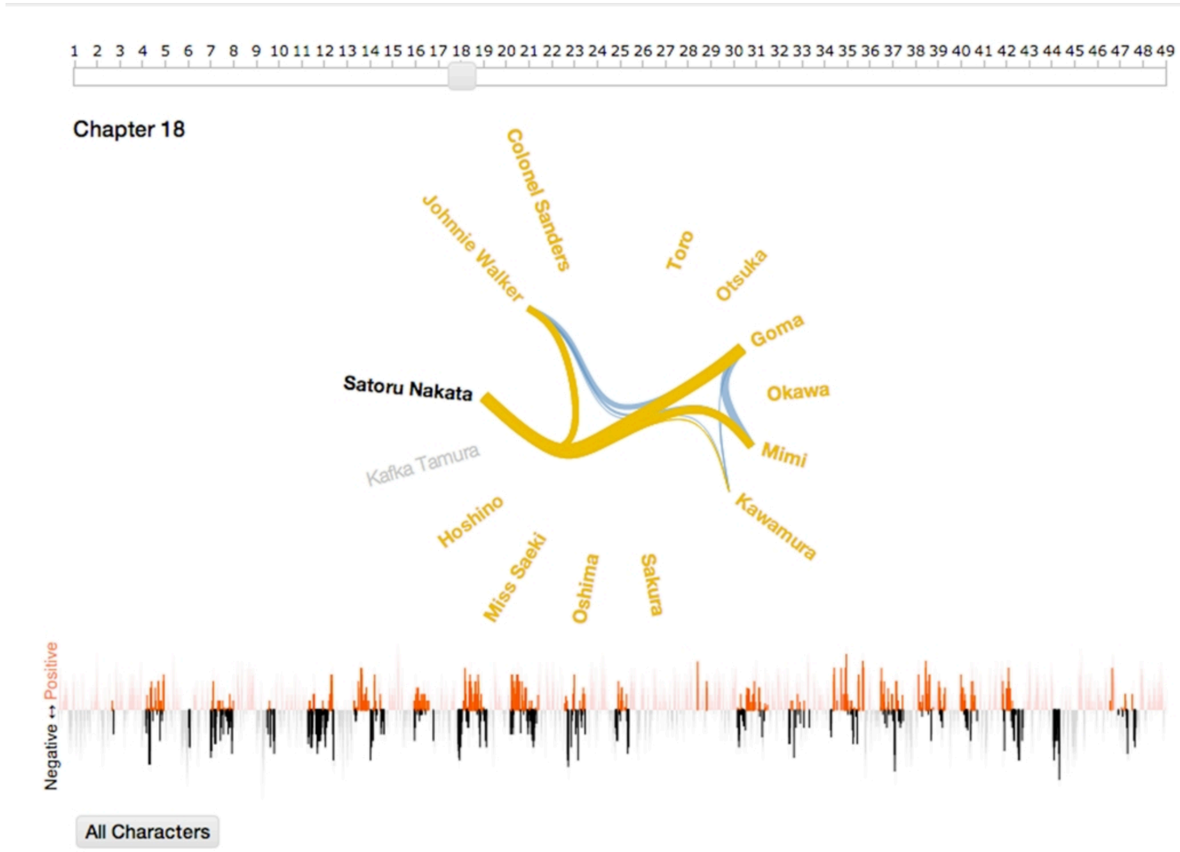
STEP 6
BEYOND WORD CLOUDS

THEMATIC FLOW OF THE DOCUMENT



STEP 6
BEYOND WORD CLOUDS

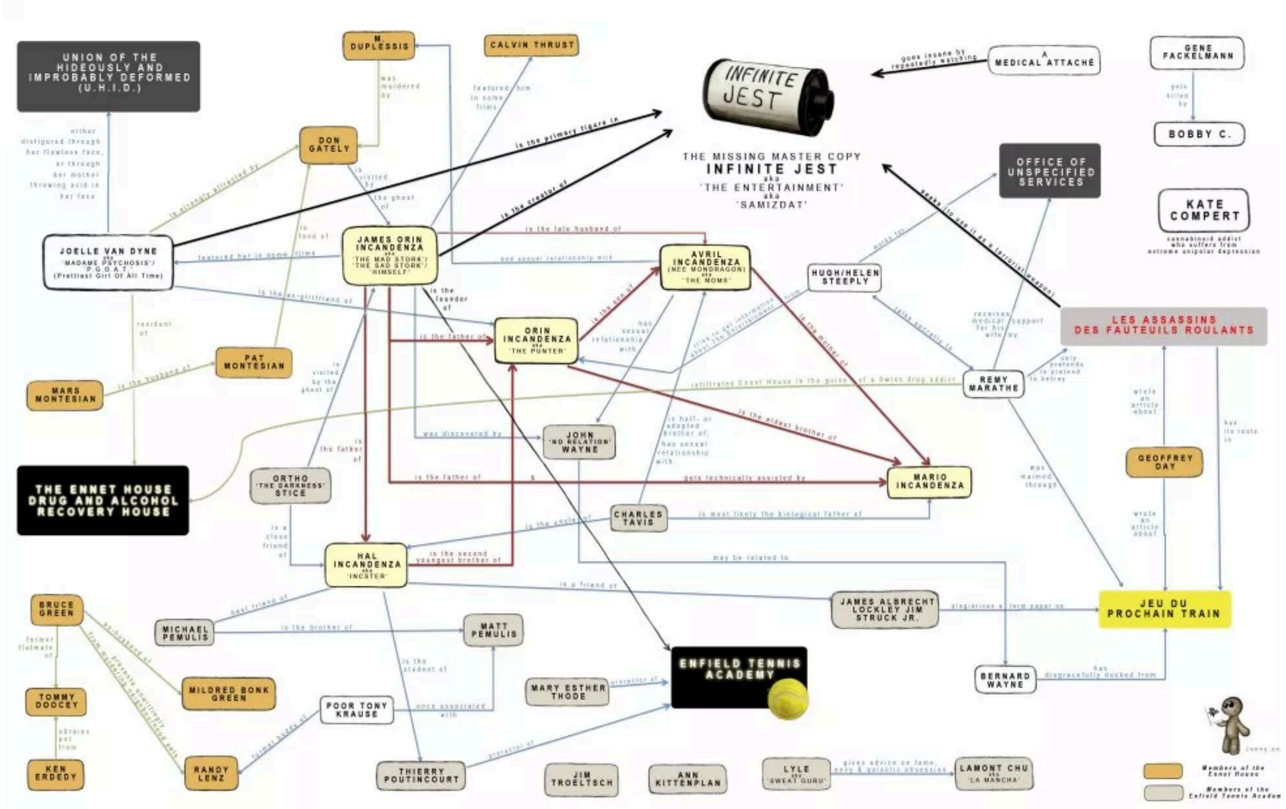
NARRATIVE LINES

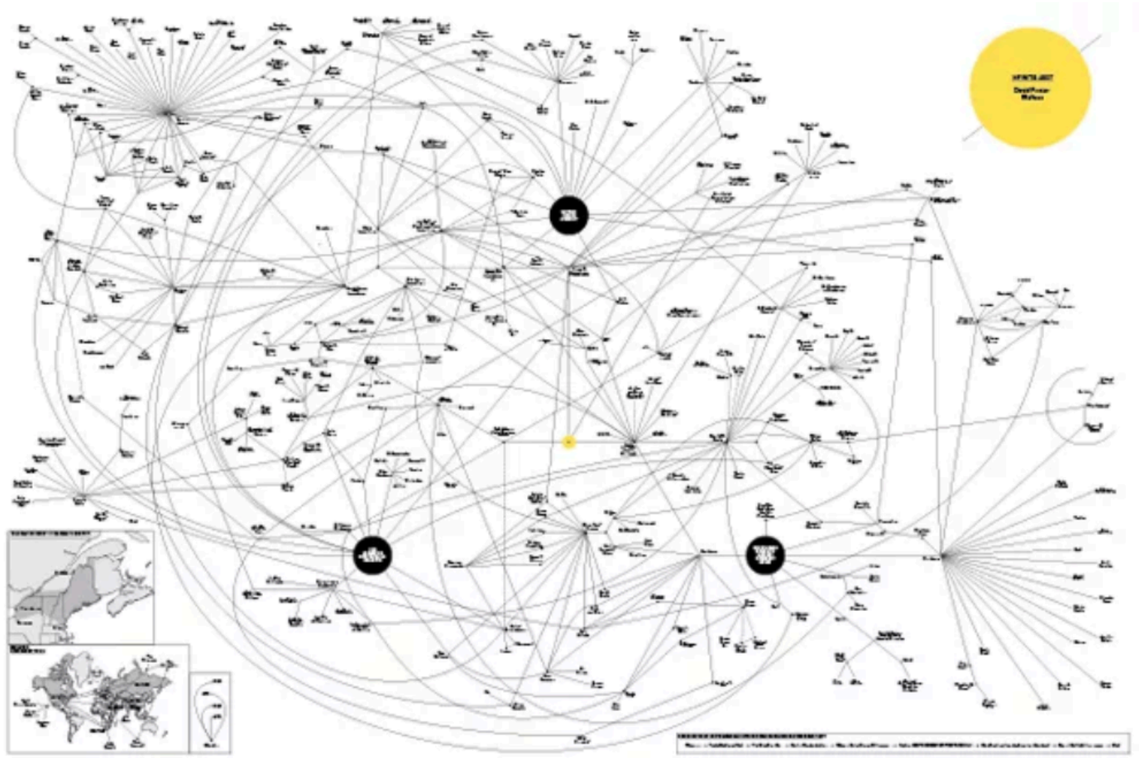


SOURCE: http://nbilenko.com/projects/narrative.html

STEP 6
BEYOND WORD CLOUDS

CONNECTION BETWEEN CHARACTERS

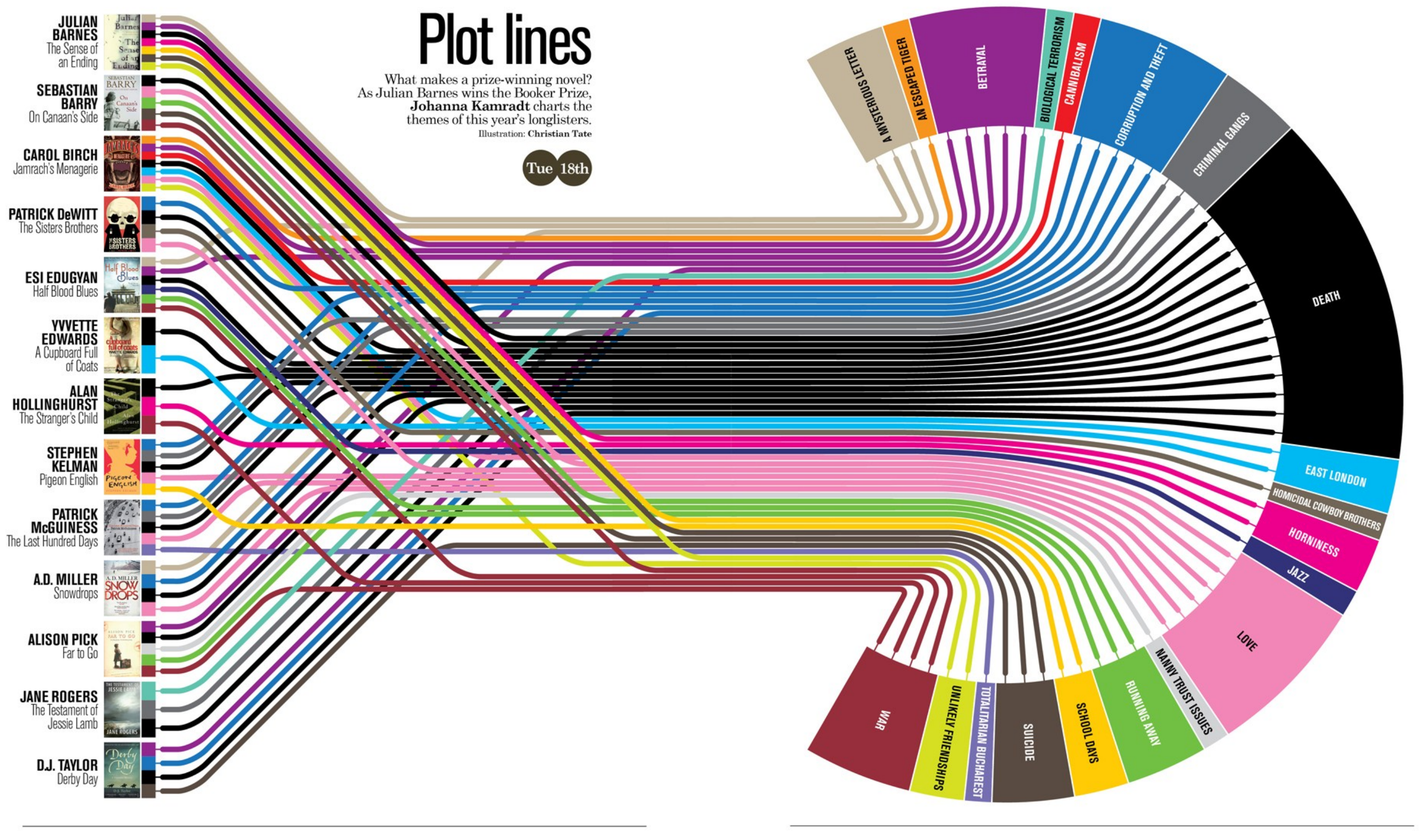




source: https://www.brainpickings.org/2010/12/10/infinite-jest-visualized/

STEP 6
BEYOND WORD CLOUDS

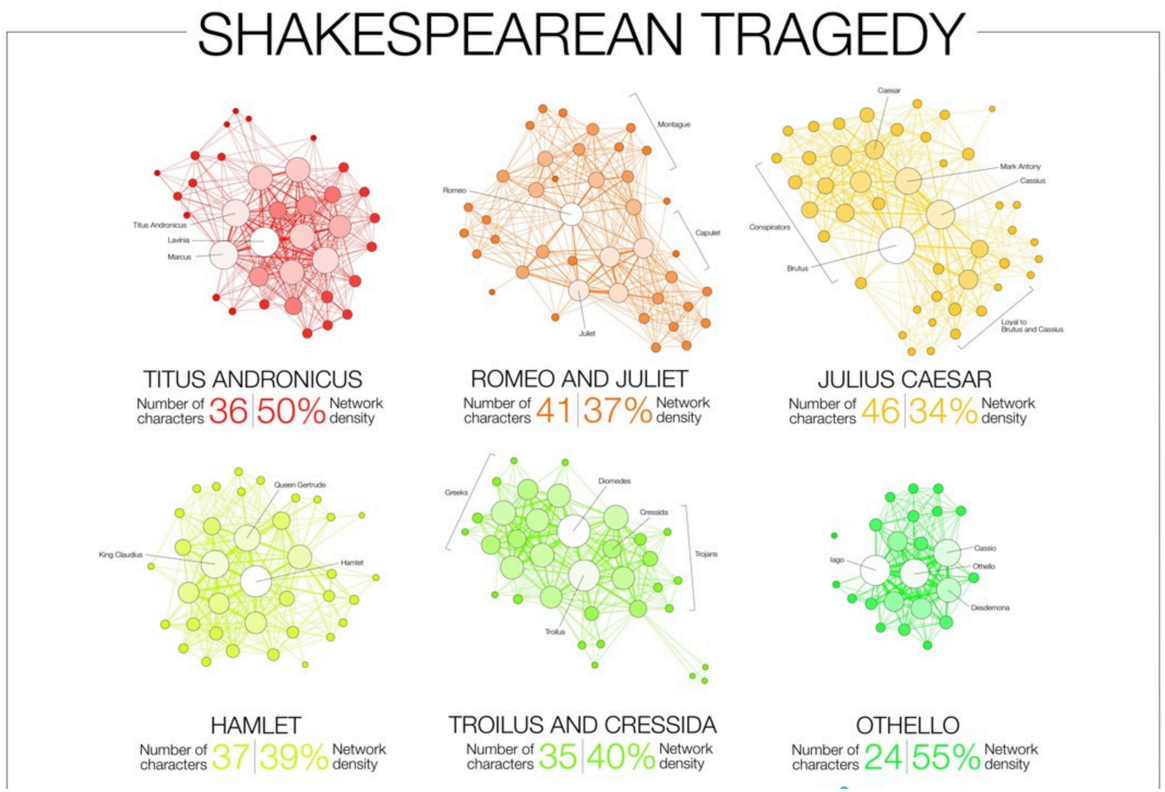
PLOT/ STORY LINES



source: https://www.brainpickings.org/2010/12/10/infinite-jest-visualized/

STEP 6
BEYOND WORD CLOUDS

CHARACTER NETWORK GRAPHS



source: https://www.brainpickings.org/2010/12/10/infinite-jest-visualized/

SUMMARY

Summary

- word clouds are bad
  - the visualization is hard to use
  - they mangle ideas by counting words
- alternative to counting words are methods that require additional text work
- Language work is still falling short in areas paraphrasing, summarization and
- There are plenty of paid APIs and free libraries in Python to perform basic ML tasks on text
- There are a lot of great visualization alternatives like narrative, story lines, but their require a lot of additional algorithmic processing
- With processes like Entity Extraction, Topic Modeling and summarization we can create better ways to highlight the ideas on unstructured text data

QUESTIONS: EMAIL ME AT ANNA.NICANOROVA AT GMAIL.COM