

WINTER INSTITUTE IN DATA SCIENCE AND BIG DATA 2023

Abdullah Yasir Atalan

1/9/23

WHAT TO COVER?

- Idea of tweet scraping and application using twarc2 in Python.
- How to analyze text as data and what can we do with the NLP?

WHY TWITTER DATA?

- It is basically public opinion.
- It may not be very representative, but it gives opportunity of a real time tracking opinions.
- Highly useful for social scientists and corporate industry. Advertisements.

TWEET SCRAPING

- Academic track in Twitter, started in 2021
- <https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api>
- Ordinary researchers 10 million tweets per month from Twitter. Business track is way more than this.

TWEET SCRAPING

- Open an academic developer account
- **Guide:**
<https://developer.twitter.com/en/products/twitter-api/academic-research/application-info>
- Specify what kind of academic research you will use it for.
- After getting authentication, now it is time to scrape tweets.
- Some tools: tweepy in Python, twarc2 in Python, twitterR in R.

TWARC2

An application using Twarc2. The logic is the same across other tools.

It is essentially a command line tool. But can be used as a Python Library.

```
pip install --upgrade twarc
```

You may need to install twarc through brew if you are on macbook

```
brew install twarc
```

Twarc will ask our tokens for authentication. BEARER TOKEN.

```
twarc2 configure
```

TWARC2

Let's start with our first search

```
twarc2 search #mush --limit 300 tweets.jsonl
```

More complex search:

```
twarc2 search --archive --limit 3000 --start-time 2022-01-01 --end-time 2022-03-01 '(#brazil OR #bolsorano OR (brazil bolsorano)) lang:en -is:retweet' tweets.json
```

- search is the tweet search command.
- --archive indicates that my developer account has academic access.

- --start-time says what date I want it to start searching for > tweets. For example, I asked it to get tweets since the new > year, 2022.

TWARC2

- --end-time tells me what date I want it to finish searching. I told > it to finish on 3rd of March. But it probably won't be able to > reach this date anyway, because I put the limit for 3000. When it > exceed 3000 tweets, it will terminate the command.
- Query.txt: Go to [Twitter Advanced > Search](#), fill > the boxes with the filters you want, then paste the query from the > search space on top-right.

TWARC2

- The tweets.json code at the end tells me with which name and > extension I will save the file. It registers in whichever > directory you are in at that moment.
- Save it into a data frame format. Twarc handles this.

```
twarc2 csv tweets.json tweets.csv
```

- You can import this dataset to your R or Python environment to do your analysis.
- Before finishing, a reminder. Do a tweet count first:

```
twarc2 searches --archive --start-time 2021-10-01 --end-time 2021-11-01 --counts-only query.txt tweet_counts.csv
```

NLP PREDICTION AND TOPIC MODELING

What is NLP?

Subfield ... concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

MAIN COMPONENTS OF AN NLP PIPELINE

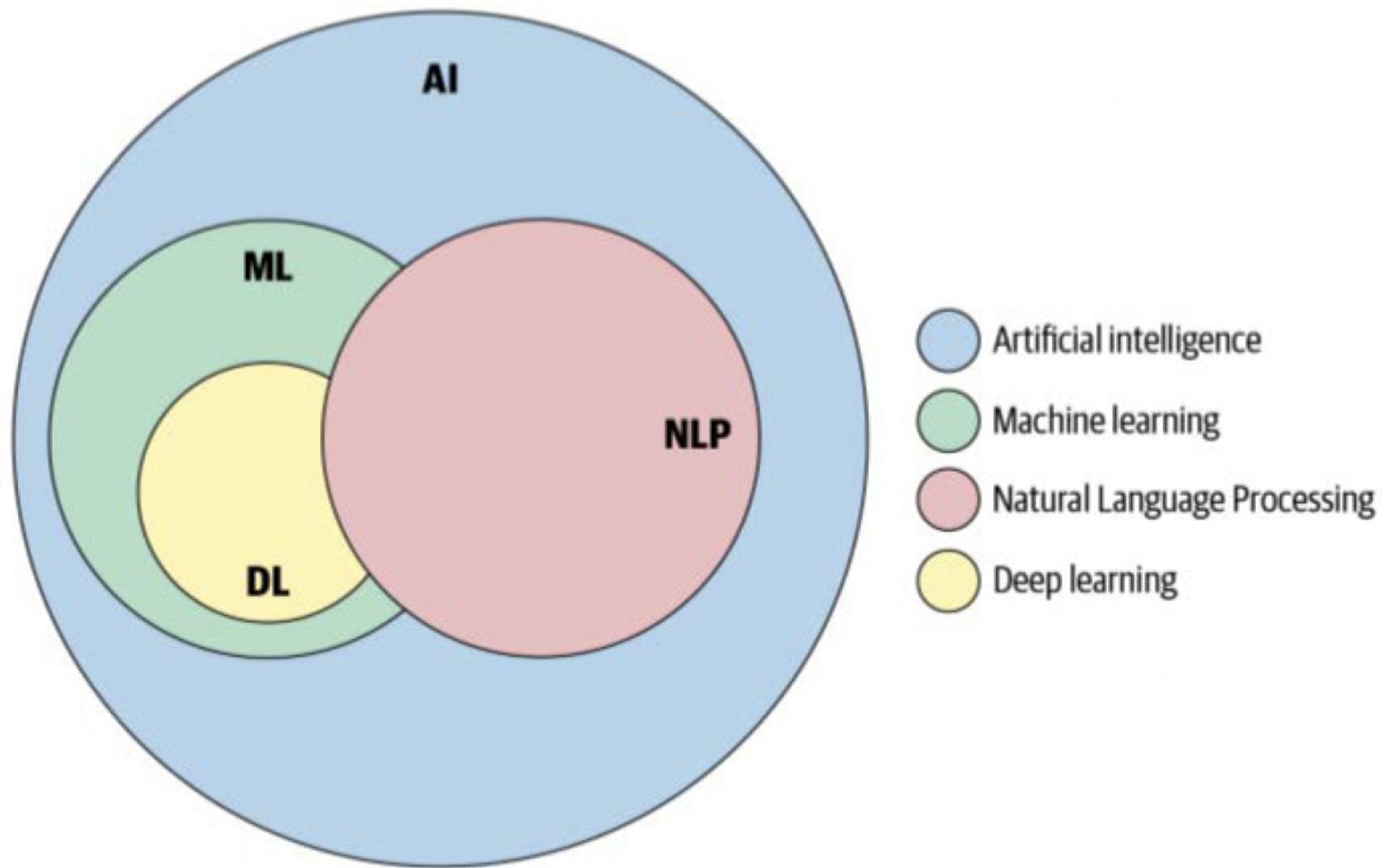
- Data Acquisition: Collect the data relevant to the given task
- Text Cleaning: Data needs to be cleaned
- Pre-processing: Convert text data into a canonical form
- Feature Engineering: Make data understandable by algorithms
- Modeling/Evaluation: Built and compare models
- Deploy model in production
- Update model to keep up model performance

NLP

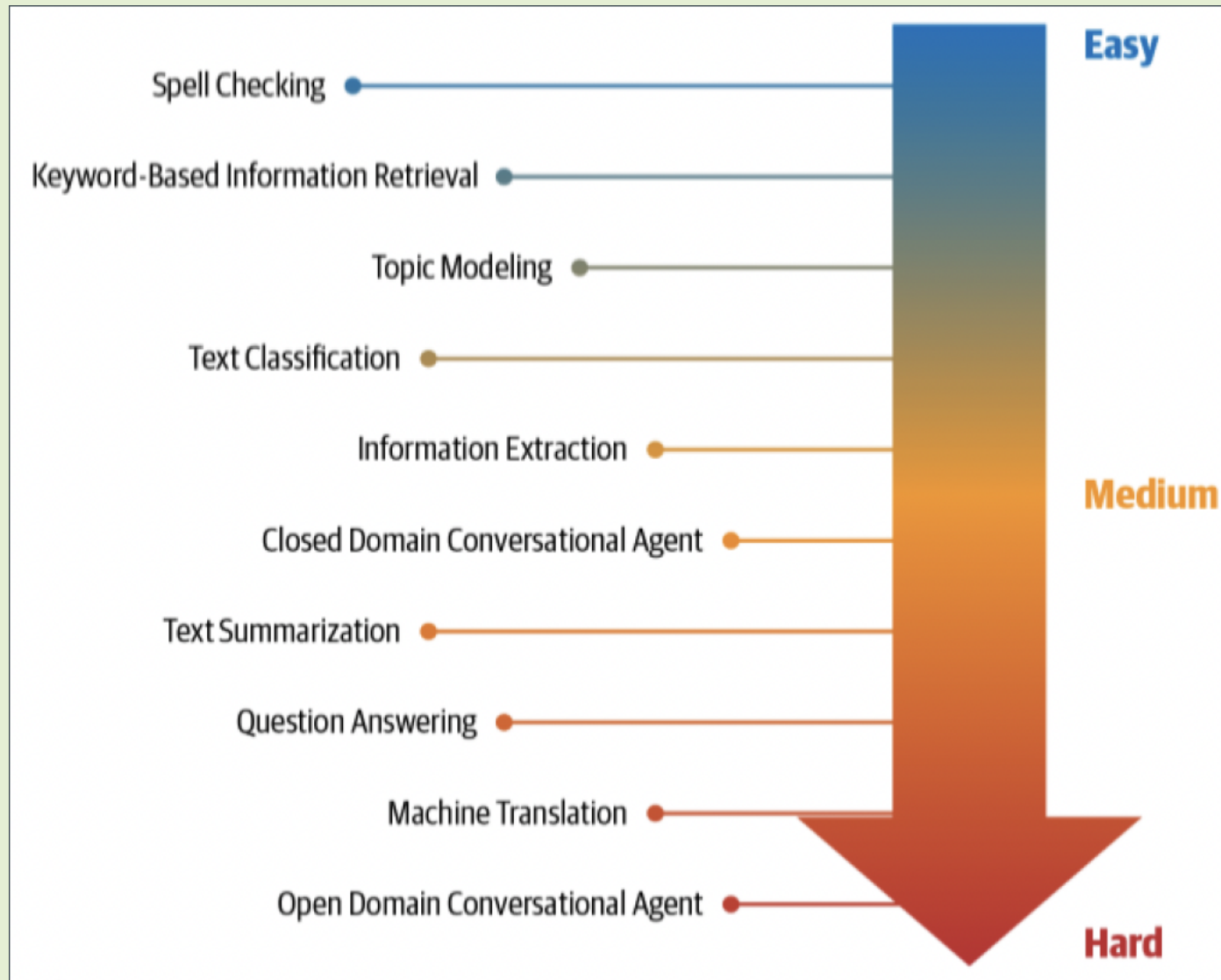
What is pre-processing?

- Removing punctuations like ., ! \$() * % @
- Removing URLs
- Removing Stop words
- Lower casing
- Tokenization
- Stemming
- Lemmatization

NLP



NLP TASKS



TOPIC MODELING

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

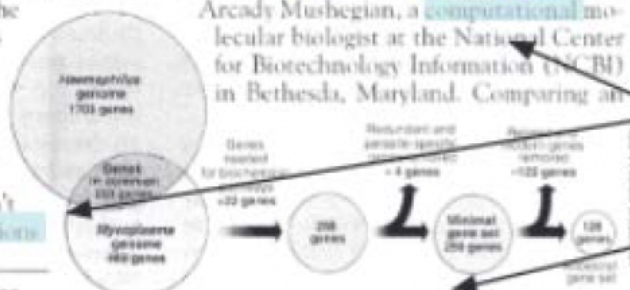
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **numeric** **numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

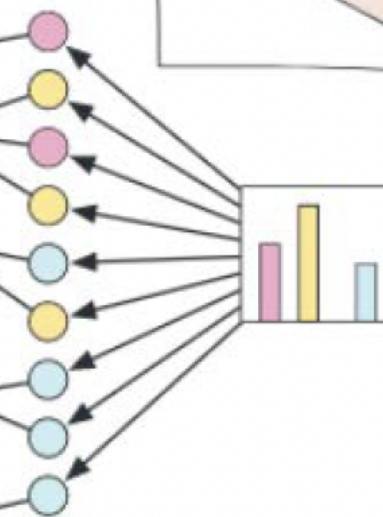


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

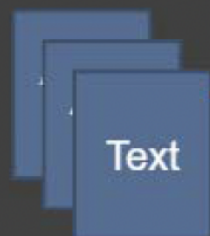
SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



TOPIC MODELING

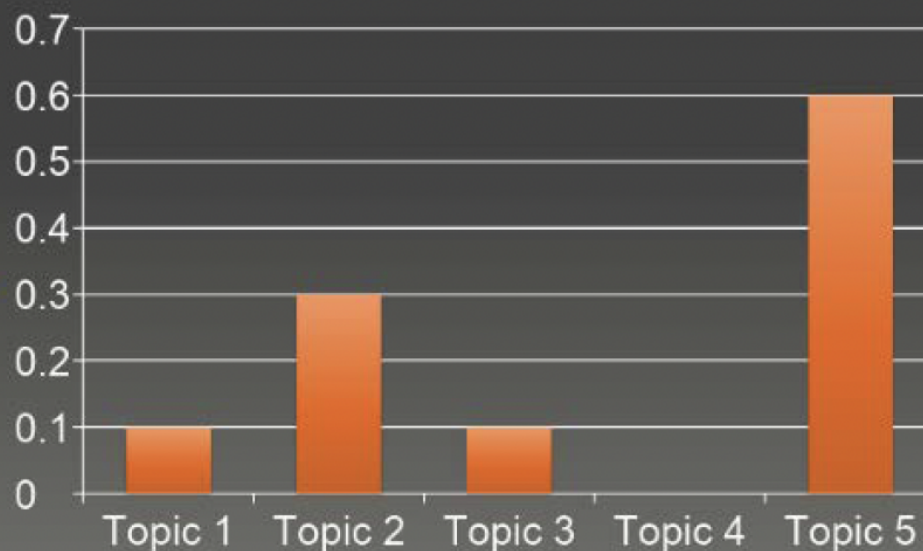
1. Identify “topics”
(groupings of words that
tend to co-occur)



Document
Topics

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|--|--|
| fees, charged, expenses, charges, paid, performance based, charge, mutual, distinct, separate, negotiable, | value, market, conditions, decline, fair, price, line, based, money, stock, relative, rates, cash, factors, valuation, | risk, loss, bear, risks, prepared, investing, involves, tolerance, liquidity, return, term, methods, market, approach, investor, | disciplinary, legal, history, criminal, regulatory, civil, report, evaluation, events, activities, personnel, material, reportable, | conflicts, potential, conflict, interests, fiduciary, arise, duty, incentive, affiliates, manner, resolve, create, avoid, |

2. Identify distribution of
topics per document



SOME TOPIC MODELING ALGORITHMS

- NMF Non-Negative Factorization
- LDA Latent Dirichlet Allocation
- SVD Singular Value Decomposition

-

