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, twitteR 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

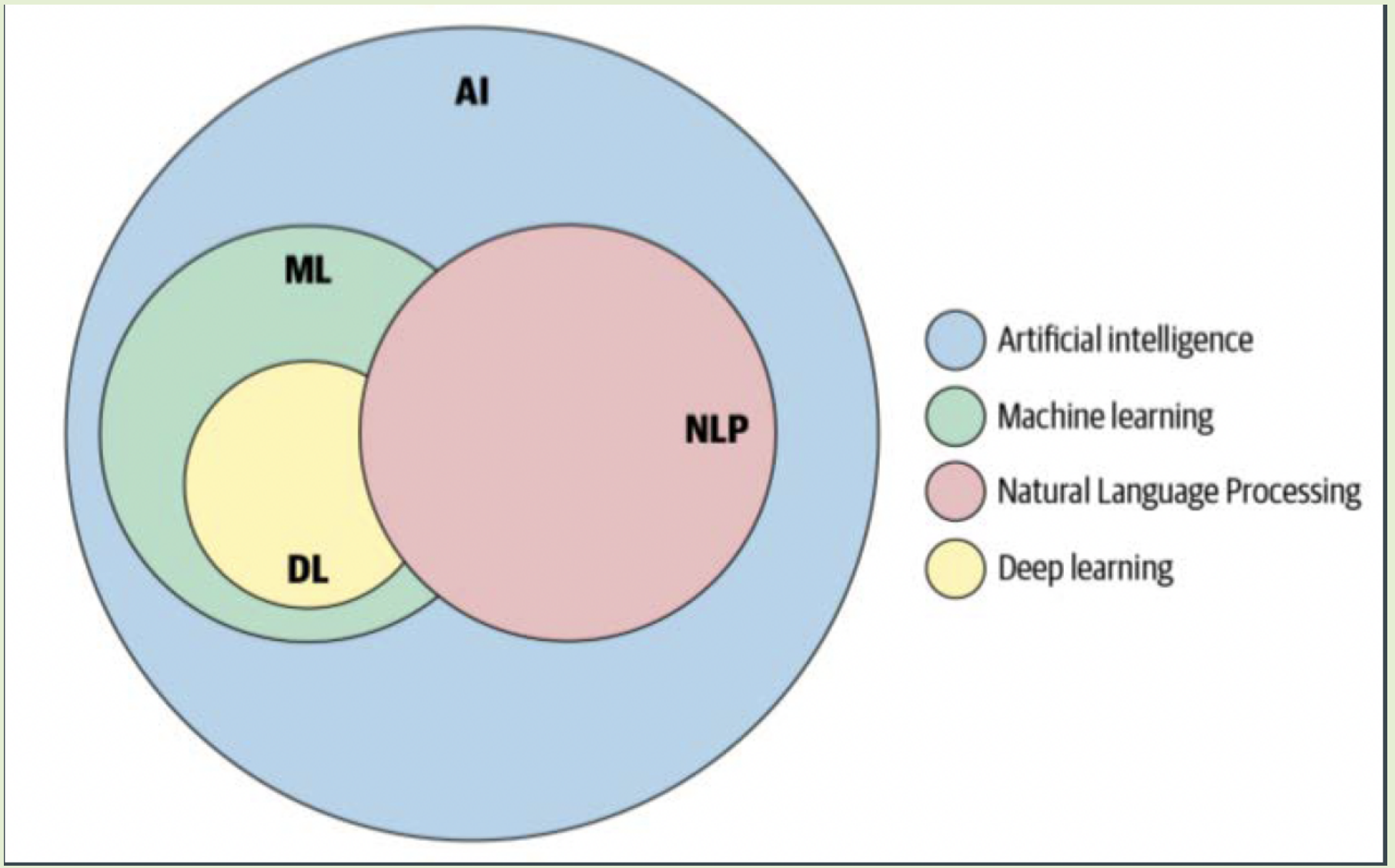
- Data Acquisition: Collect the data relevant to the given task
- Text Cleaning: Data needs to be cleaned
- Pre-processing: Convert text data into a canonical form
- Feature Engineering: Make data understandable by algorithms
- Modeling/Evaluation: Built and compare models
- Deploy model in production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

NLP

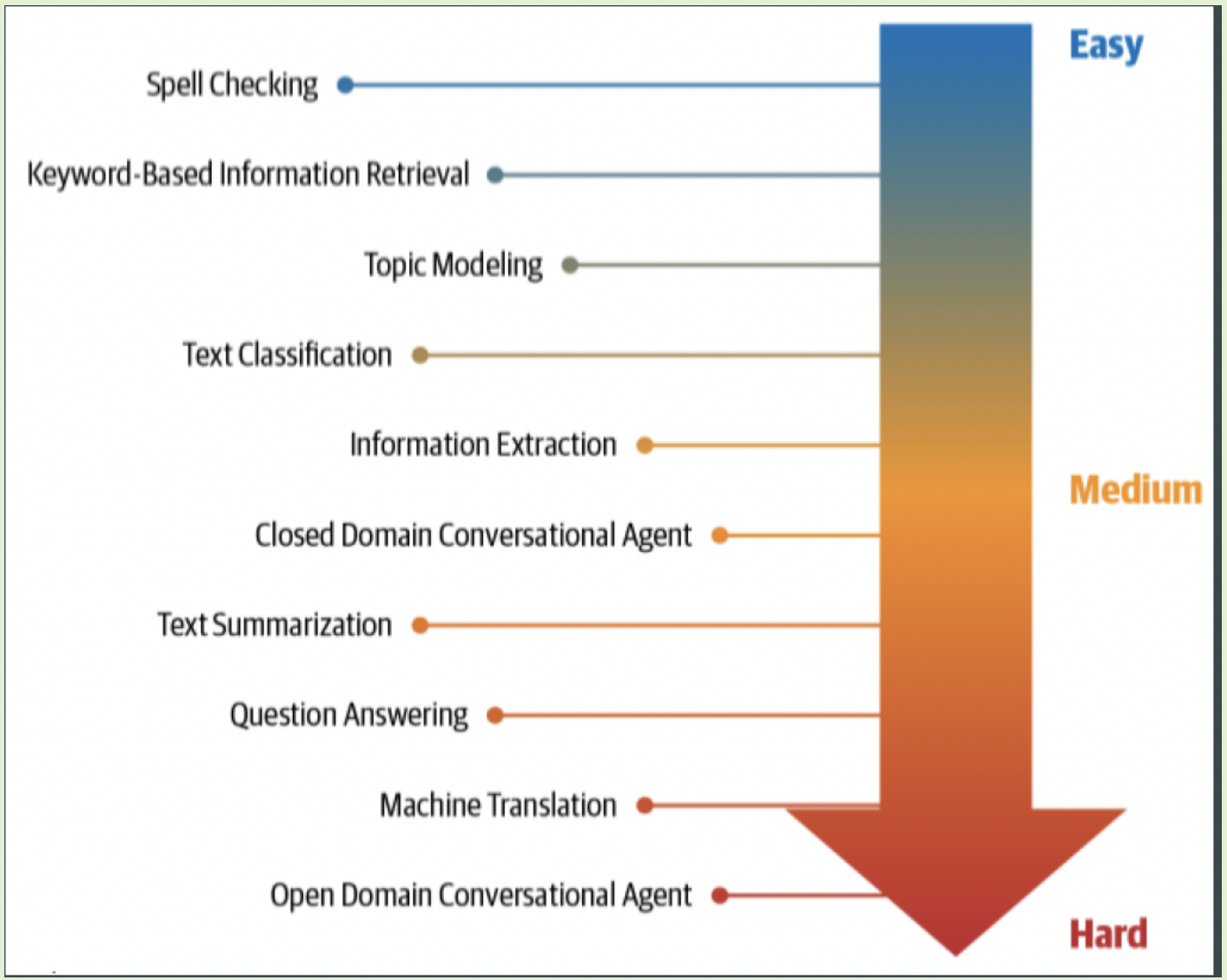
What is pre-processing?

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

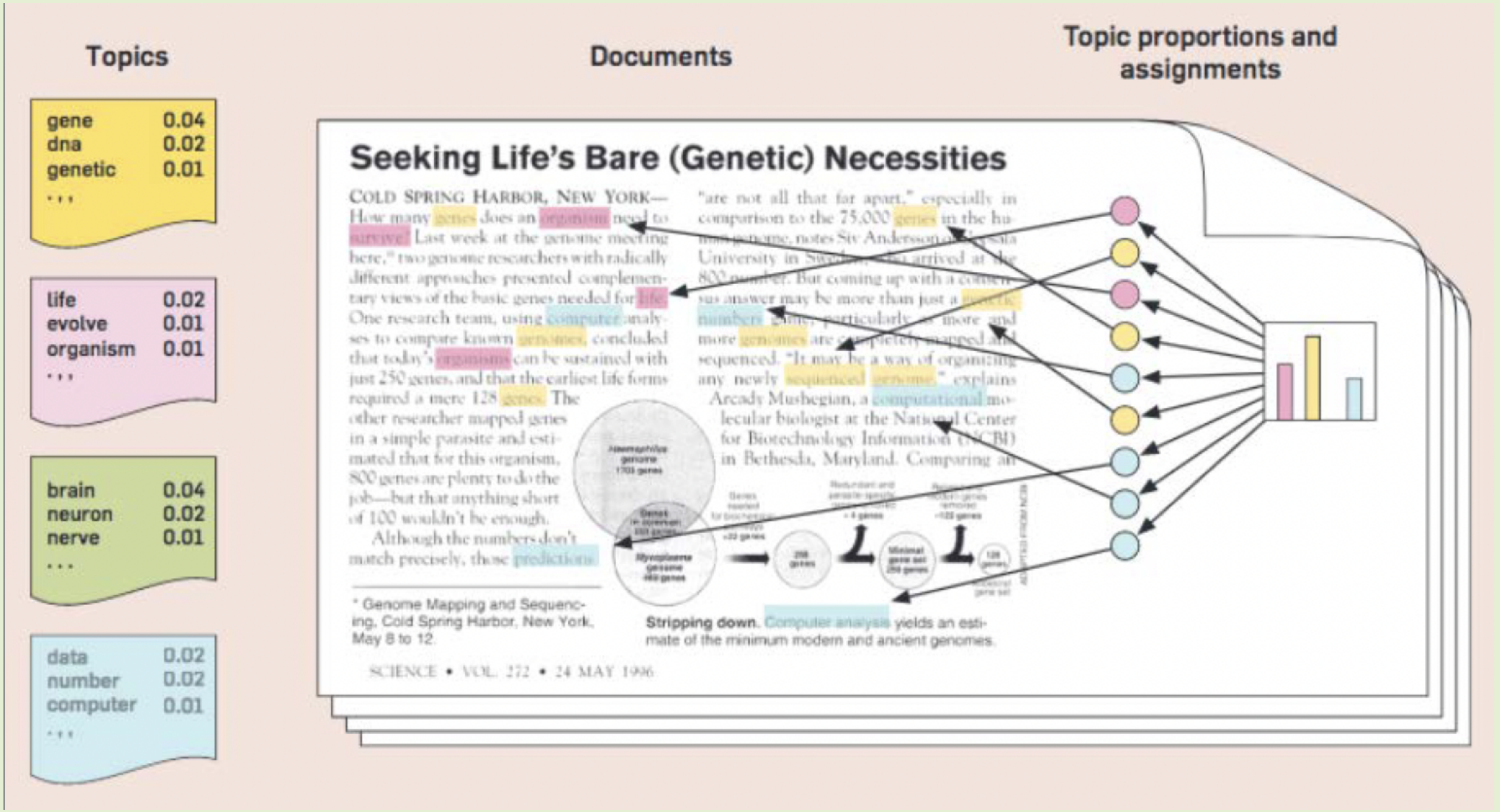
NLP


AI
ML
NLP
DL
Artificial intelligence
Machine learning
Natural Language Processing
Deep learning

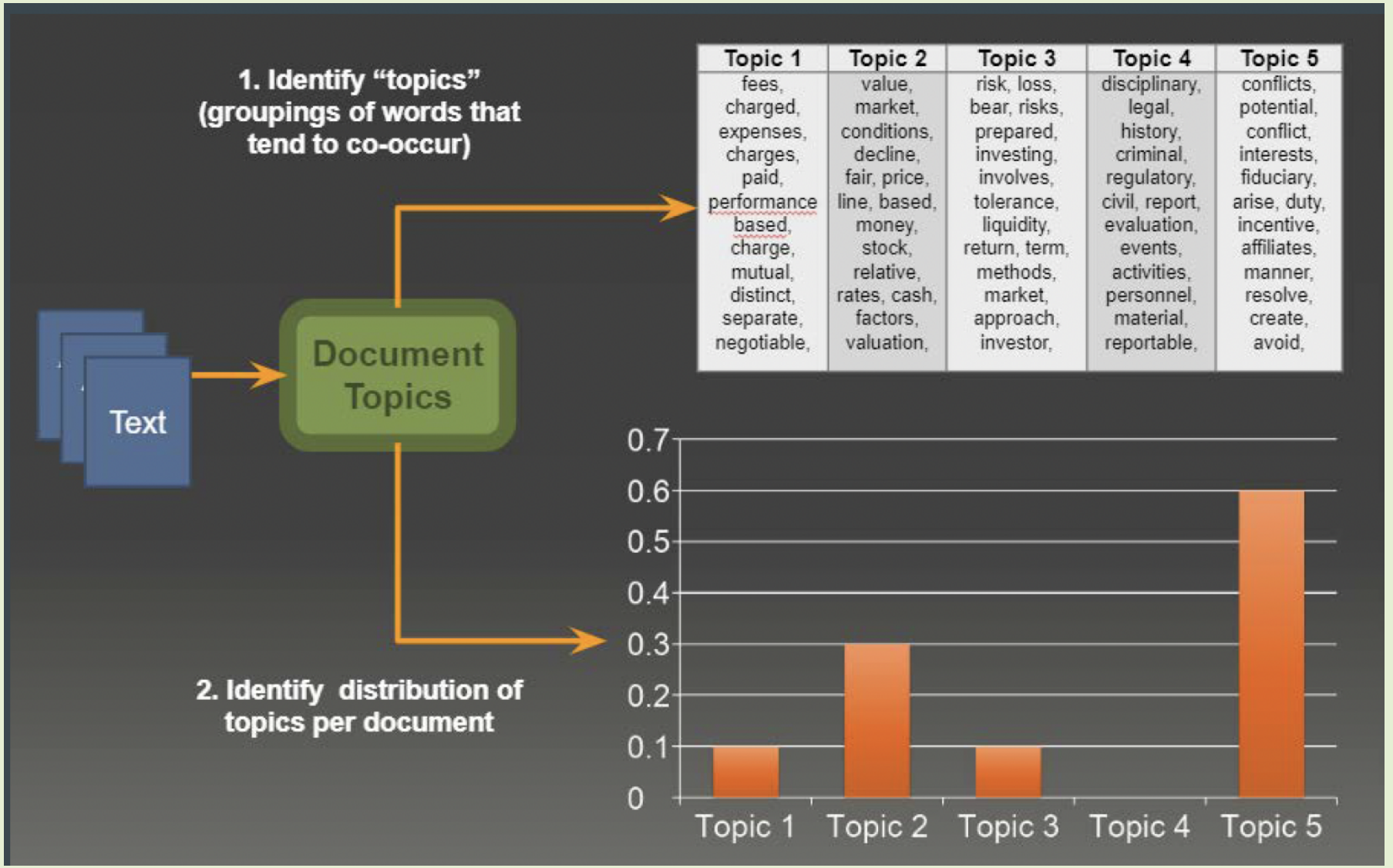
NLP TASKS


Spell Checking
Keyword-Based Information Retrieval
Topic Modeling
Text Classification
Information Extraction
Closed Domain Conversational Agent
Text Summarization
Question Answering
Machine Translation
Open Domain Conversational Agent
Easy
Medium
Hard

TOPIC MODELING



TOPIC MODELING



SOME TOPIC MODELING ALGORITHMS

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