Understanding Texts via Topic Extraction (an introduction)

David Przybilla

@dav009

http://github.com/dav009

http://alejandro.pictures

idio
BIG DATA COLOMBIA
DBpedia
ukgovcamp
OPEN DATA

Notebook + Slides

https://github.com/dav009/topictalk

hope you pulled the docker image at home

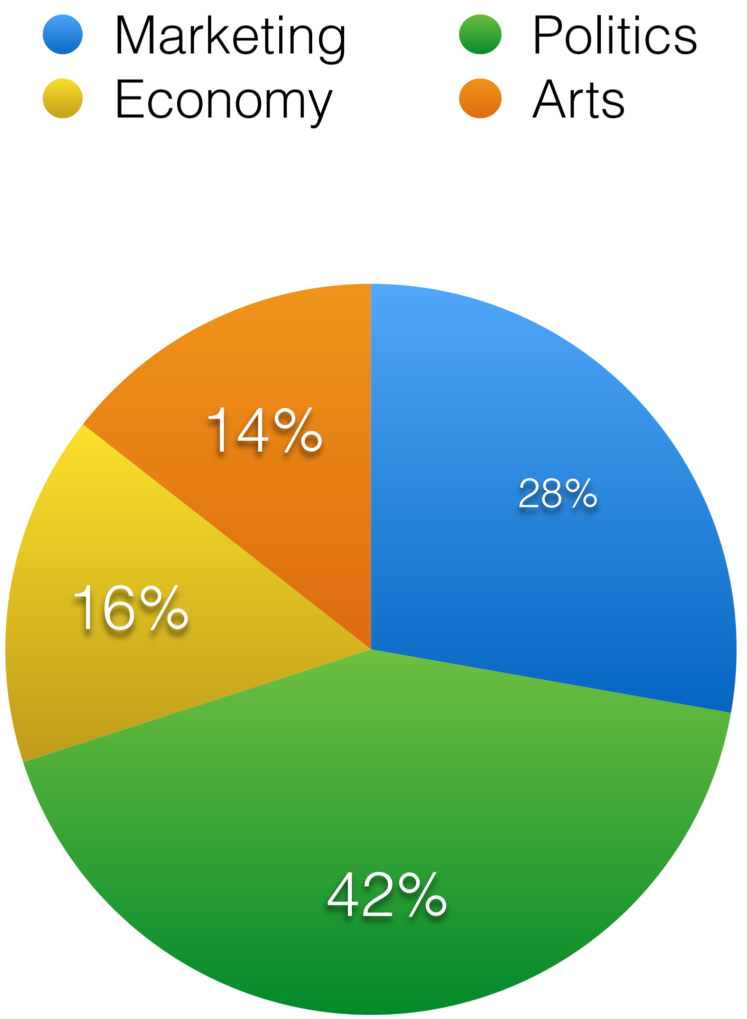
I struggled defining the scope of this talk 😓

feel happy to poke me if you want to:

- hack
- bring the discussion forward

Companies have **lots** of **unstructured data**

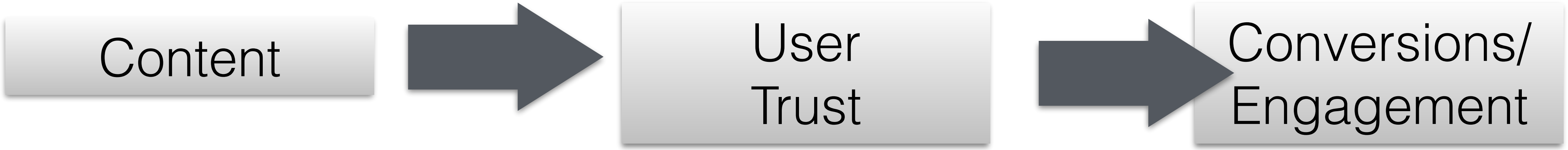
Topics can give a high overview of a companies assets



Why ?

Actionable Metrics

Buying / Creating content (Content Marketing)



Content

User Trust

Conversions/ Engagement

- What content is generating me conversions?

- What kind of content generates me **more** **conversions/engagement**?

Topics!

lots of knowledge about what they do

Search problem

Search Keyword: "*Video game*"

Document: "*Mario, Bros……,Nintendo*"
(No mention of Video game)

Feedback from users

E-commerce

- Users are talking more this month about *Swimwear* than previous months
  - Maybe we should publish more products in that category?

Understandable Profiles

```
Client: {
    Manchester United: 0.8
    Football: 0.9
    sushi: 0.1
}
```

- Client: I want to quit your service!
- Customer service: stay, we give you free *Manchester united* games for the rest of the year 😏
- Client : yay 😗

get the **gist** of lots of text

- Wikileaks
- Panama papers

Topics!

- Use them as features for other tasks:
    - Recommendation Systems
    - Creating profiles of Users that are ‘Understandable’
    - Information Retrieval: Better Search results
    - Question & Answer
    - Semi automatic Ontology creation

How?

- Classic Name Entity Recognition (NER)
- Name Entity Linking (NEL)
- Topic Modelling (Too much here…)

Name Entity Recognition

Barack Obama, Person

Vietnam, place

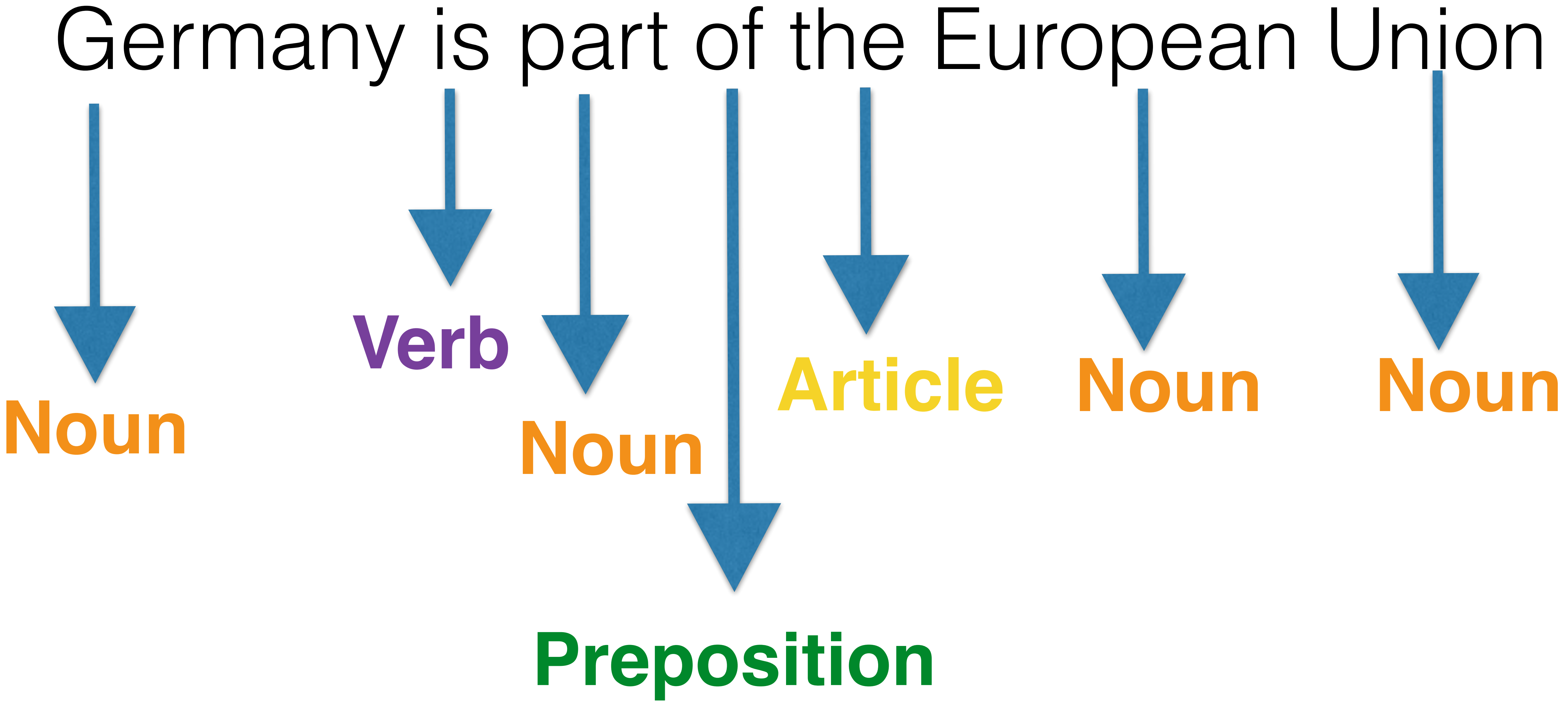
Apple Inc, Organisation

Traditional Task only three categories :
Person, place, Organisation

Name Entity Recognition

1. Part Of Speech tagging
2. Linguistics Rules

PoS (part of speech Tagging)



Germany is part of the European Union

Noun

Verb

Noun

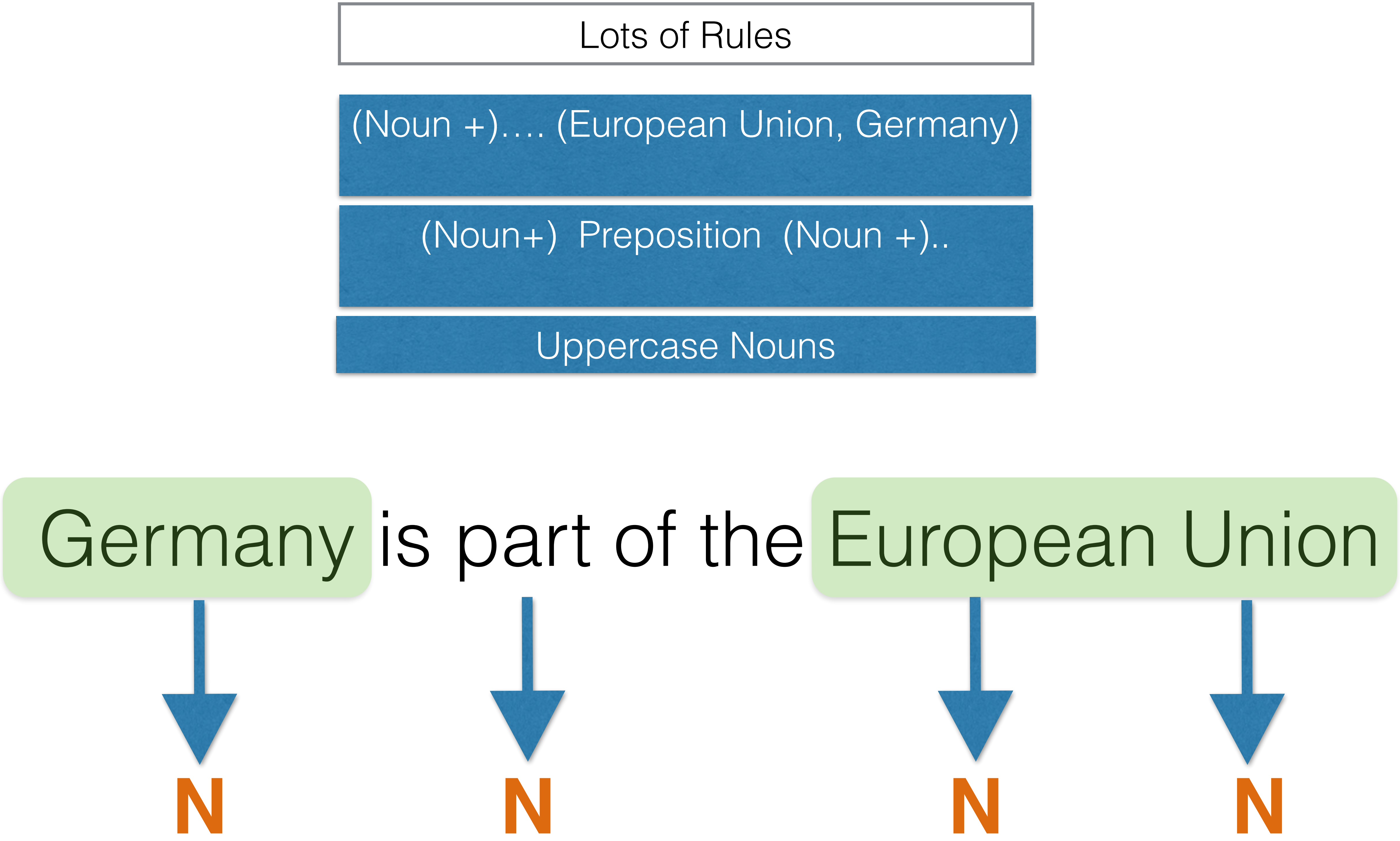
Preposition

Article

Noun

Noun

Name Entity Recognition



Lots of Rules

(Noun +)…. (European Union, Germany)

(Noun+) Preposition (Noun +)..

Uppercase Nouns

Germany is part of the European Union

Name Entity Recognition

The **bad** 😒:

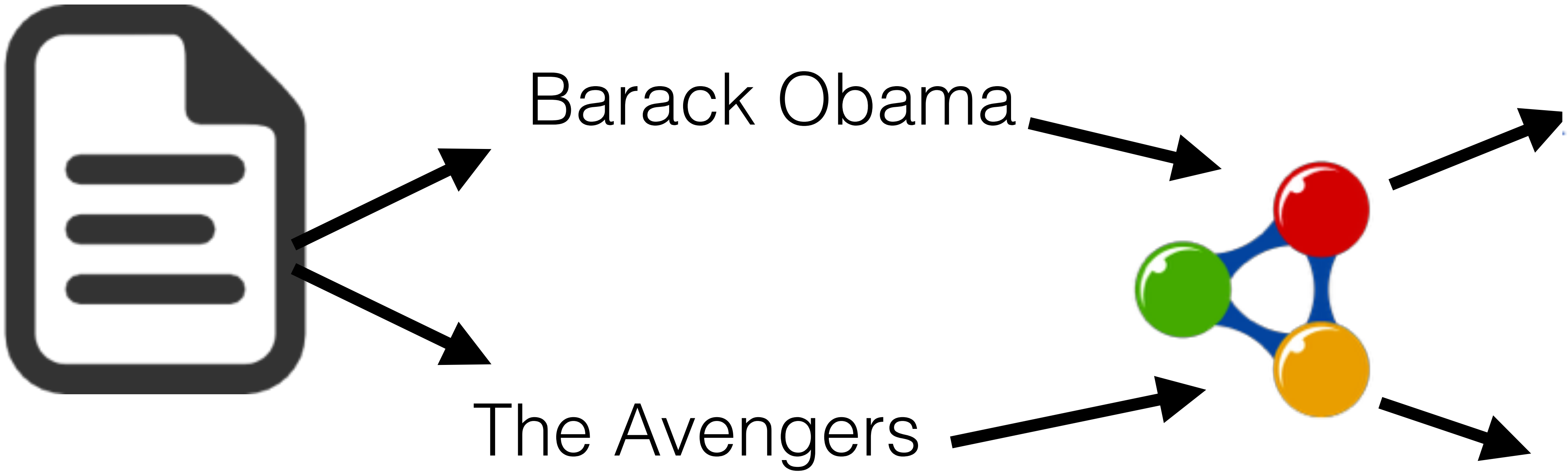
- You have to define lots of rules
- Dependent on the quality of PoS
    - noisy text i.e: capitalisations are wrong

Name Entity Recognition

- The **good** 🙃:
  - Lots of packages do it out of the box (i.e: NLTK)
  - German, Spanish, French…
  - No idea about Vietnamese ???

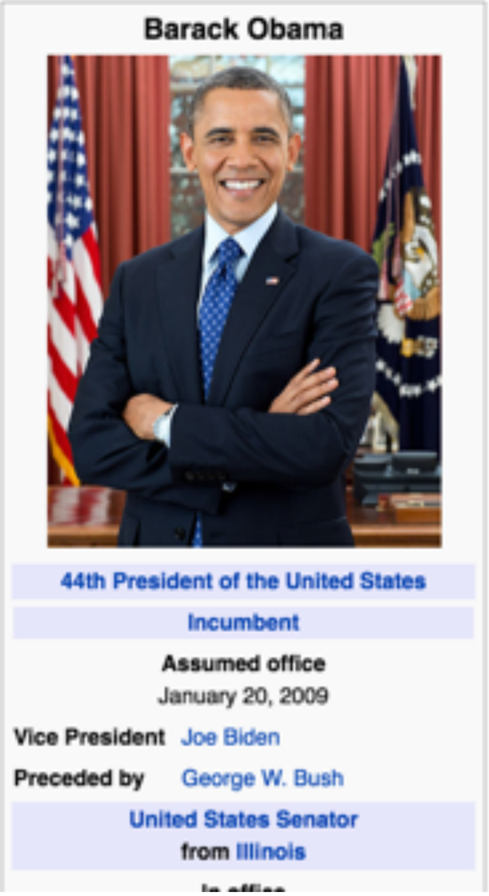
Notebook time! 😗

Entity Linking



Barack Obama

The Avengers



Barack Obama

44th President of the United States

Incumbent

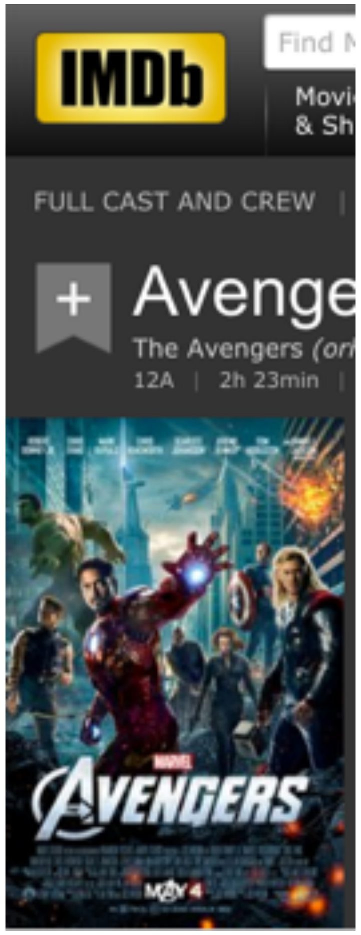
Assumed office
January 20, 2009

Vice President Joe Biden

Preceded by George W. Bush

United States Senator
from Illinois

In office



IMDb

Find M

Movi
& Sh

FULL CAST AND CREW |

Avenge

The Avengers (ori

12A | 2h 23min |

🤔

Wait! what’s the difference?!

Entity Recognition: You get strings back

“Barack Obama”

“Obama”

“B. Obama”

“New York City”

“NYC”

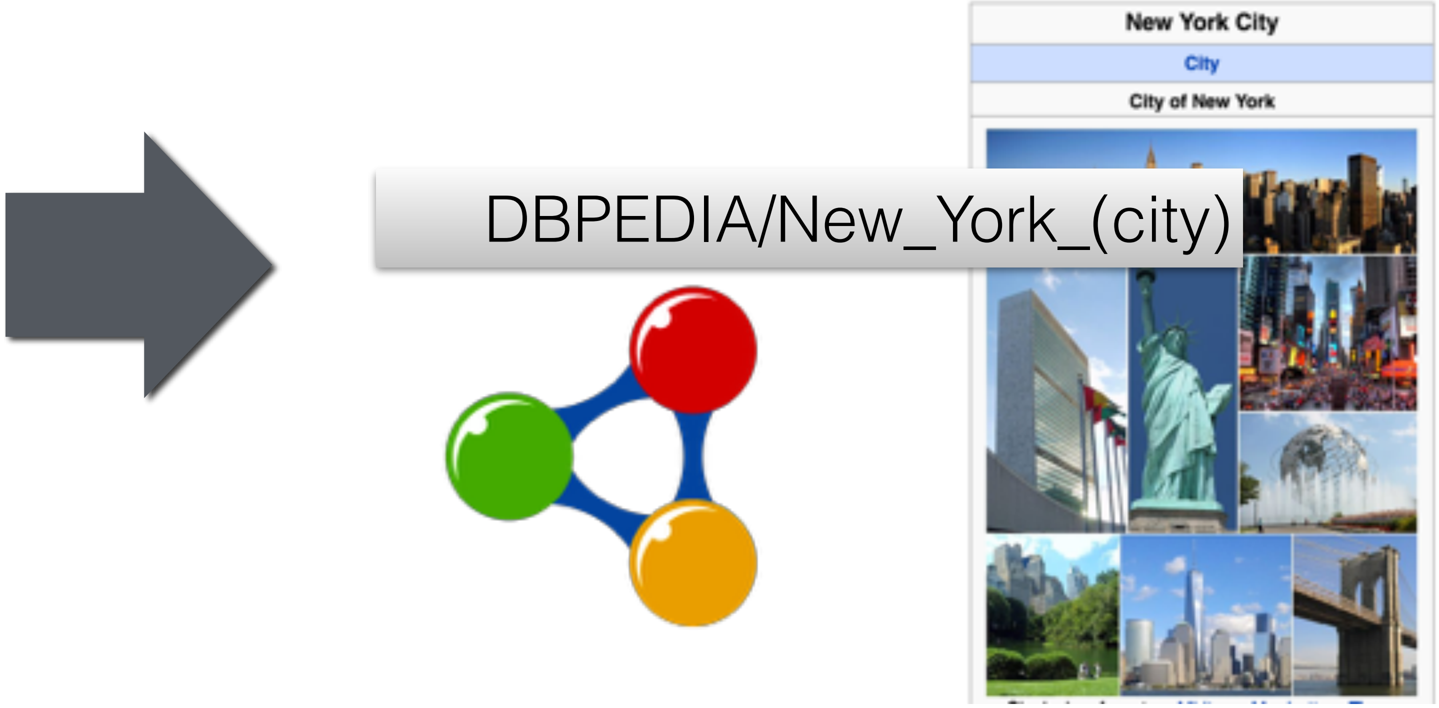
“The Big Apple”

Entity Linking: You get an Identifier back.

“New York City”

“NYC”

“The Big Apple”



DBPEDIA/New_York_(city)

DBPEDIA/New_York_(city)

New York City

City

City of New York

City in USA

Landmarks

….

Ontologies / Knowledge Bases

Freebase™ 🤕 →

WIKIDATA

DBpedia

Entity Linking

1. Find text entailing entities

2. Choose the right entity

Find text entailing entities

Apple released a new iPhone

Apple released a new iPhone

Apple: 🍎,

iPhone: 📱

iPhone: 📱

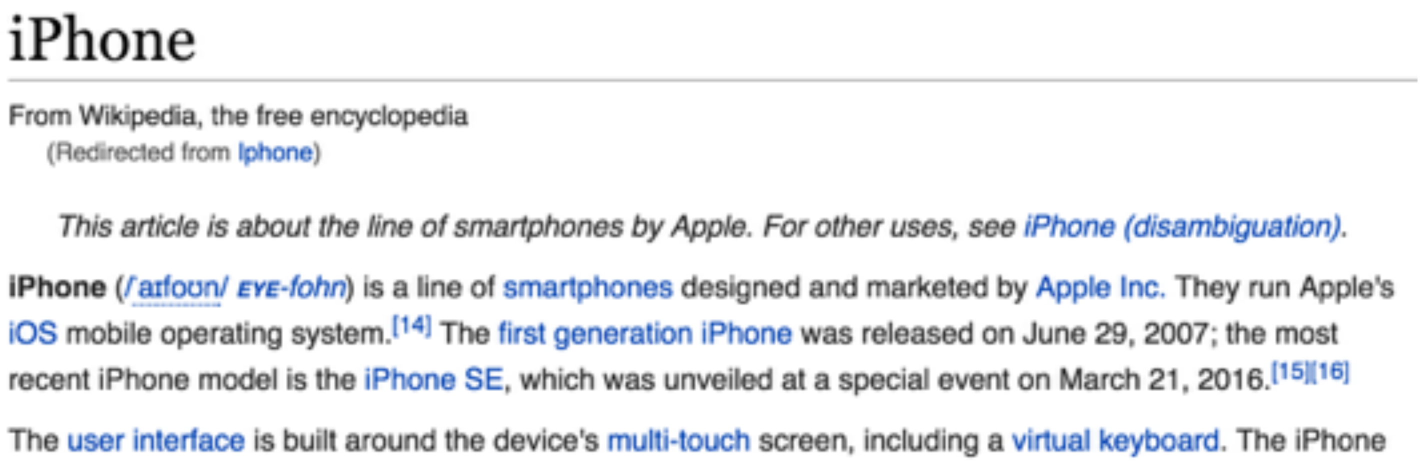
Apple: 🍎, →

disambiguate(context,🍎,)

[iphone, mac, software..] →

[fruit, flavour, colour…] → 🍎

Magic ?



iPhone

From Wikipedia, the free encyclopedia
(Redirected from Iphone)

This article is about the line of smartphones by Apple. For other uses, see iPhone (disambiguation).

iPhone (/ˈaɪfoʊn/ *EYE-fohn*) is a line of smartphones designed and marketed by Apple Inc. They run Apple's iOS mobile operating system.[14] The first generation iPhone was released on June 29, 2007; the most recent iPhone model is the iPhone SE, which was unveiled at a special event on March 21, 2016.[15][16]

The user interface is built around the device's multi-touch screen, including a virtual keyboard. The iPhone

WIKIPEDIA
The Free Encyclopedia

- “Apple” => [Apple_Inc]
- “Obama” => [Barack_Obama]
- “B. Obama” => [Barack_Obama]

The **bad** 😒:

- No links in Wikipedia
- Very Specific Domains

- The **good** 🙃:
    - Ready to use out of the box
    - It can be easily tuned adapted
    - Information about topics can be expanded
        - Via Ontology
        - Via Embeddings

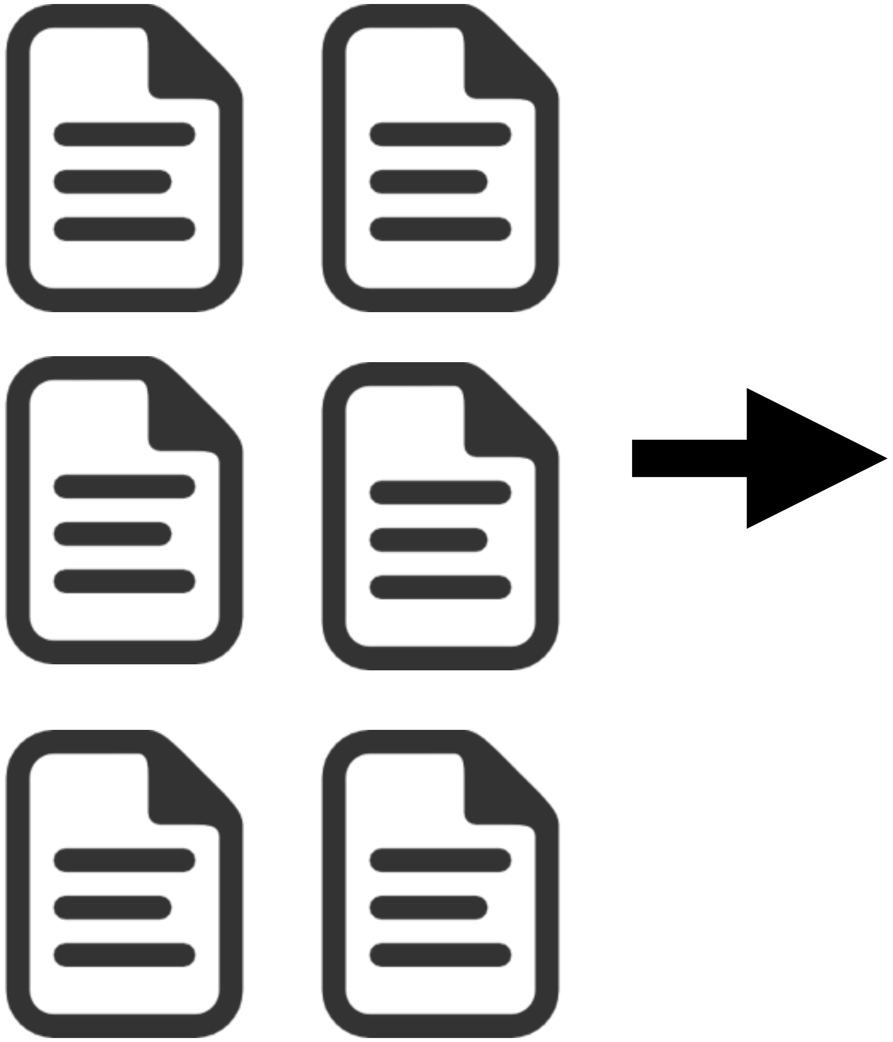
notebook time!

(a bit of cheating this time 😧)

Topic Modelling (LDA)

Latent Dirichlet Allocation

Topic Modelling (LDA)



Topic 1: [Barack, Obama, USA, President]

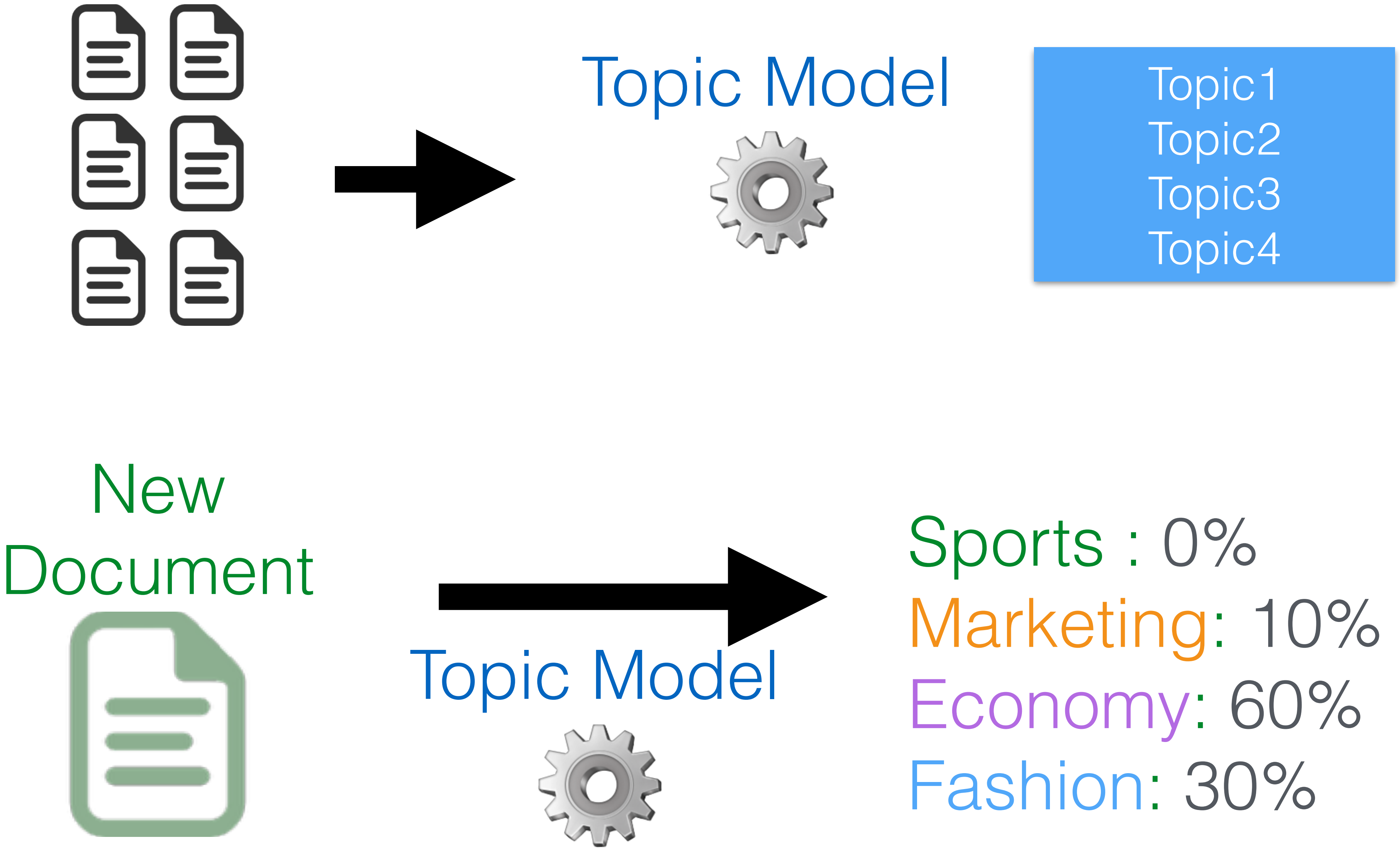
Topic 2: [Vietnam, Saigon, Pho…]

Topic N: [Sushi, fish, sea…..]

Whats the difference?

- Name Entity Recognition
  - Topic: “Barack Obama”(Just a String)
- Name Entity Linking
  - Topic: “Wikipedia/Barack_Obama”
- LDA
  - Topic: [Word1, Word2..WordN]

Topic Modelling (LDA)



Topic Model

Topic1
Topic2
Topic3
Topic4

New Document

Topic Model

Sports : 0%
Marketing: 10%
Economy: 60%
Fashion: 30%

LDA

- Assumes a document follows a discourse about topics

- Assumes words appearing in the same document are related

LDA

- Guess the number of Topics (n)

- Start by assigning Words to random topics

Topic 1

- Health
- Disease
- Cancer
- Operation

..

..

- Nintendo

Topic 2

- Games
- Playstation
- Sony
- Portable

..

..

- Nintendo

How to decide to which topic the word “Nintendo” belongs?

- A word can belong to more than one Topic

Document

does Nintendo belong to Topic1 or Topic2 in this Document?

- How often “Nintendo” occurs in documents talking about Topic Z

- How common(likely) is Topic Z for the given document

$$P(Z|W,D) = \frac{\#\ of\ word\ W\ in\ topic\ Z + \beta_w}{total\ tokens\ in\ Z + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha)$$

LDA

- Iterate X times through all the Words & Documents

- Model’s perplexity will decrease
  (the model fits the training corpus)

Notebook time!

⏱️

The **bad** 😒:

- Guessing the number of topics
- There are other parameters to guess (Alpha, Beta)
- Evaluating them (recently cool metrics have been introduced)
- Garbage Topics

- The **good** 🙃:
    - Unsupervised (No need to process wiki, no annotation)

For the curious

- LDA2VEC (Word embeddings + LDA) :)
- Evaluation (😕)

Thank you