Workshop on Topic Modeling

Jonas Frankemölle

CDHU

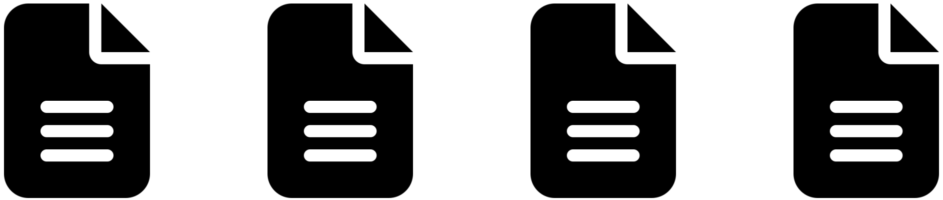
UPPSALA
UNIVERSITET

Agenda

1. Introduction of Topic Modeling and Latent Dirichlet Allocation (LDA)
2. Demo in Python
3. Hands-on with Topic Modeling

Documents

Doc 1 Doc 2 Doc 3 Doc 4



Words

Topics Politics Sports Politics Medicine

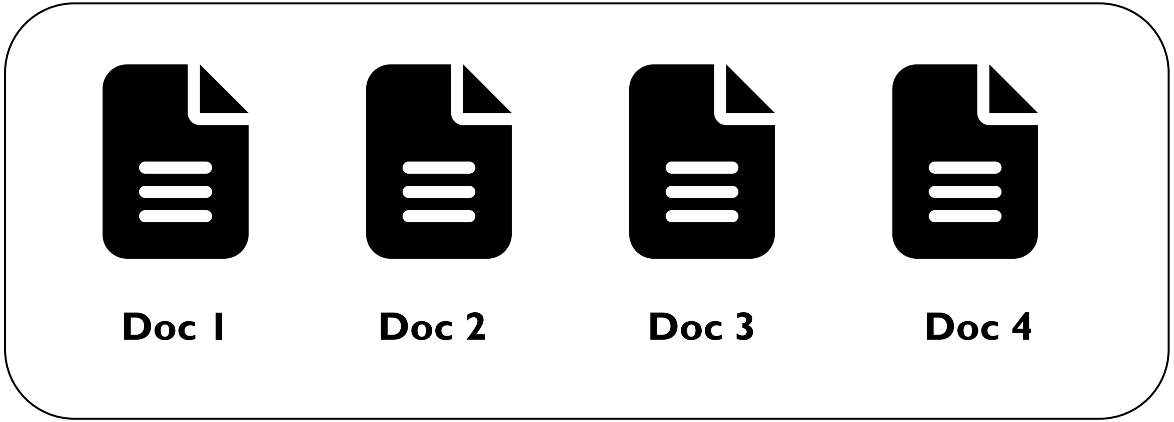
Sports

Goal is to assign topics to documents, without knowing the topics

We want a fast, unbiased way to find topics in potentially large documents

→ Topic Modeling

Corpus



Corpus: Collection of Documents

Documents

Words

Topics Politics | Sports | Politics | Medicine, Sports

Corpus: Collection of Documents

Document: Collection of topics

CDHU

UPPSALA UNIVERSITET

Politics

Election
Government
Debate
Law

Corpus: Collection of Documents

Document: Collection of topics

Topic is a collection of words ("keywords"). By looking at the words in a topic, one can identify what the topic is about

Latent Dirichlet Allocation (LDA)

- “Unsupervised Learning”
- The algorithm details are a bit complicated, but I’ll give you an intuition

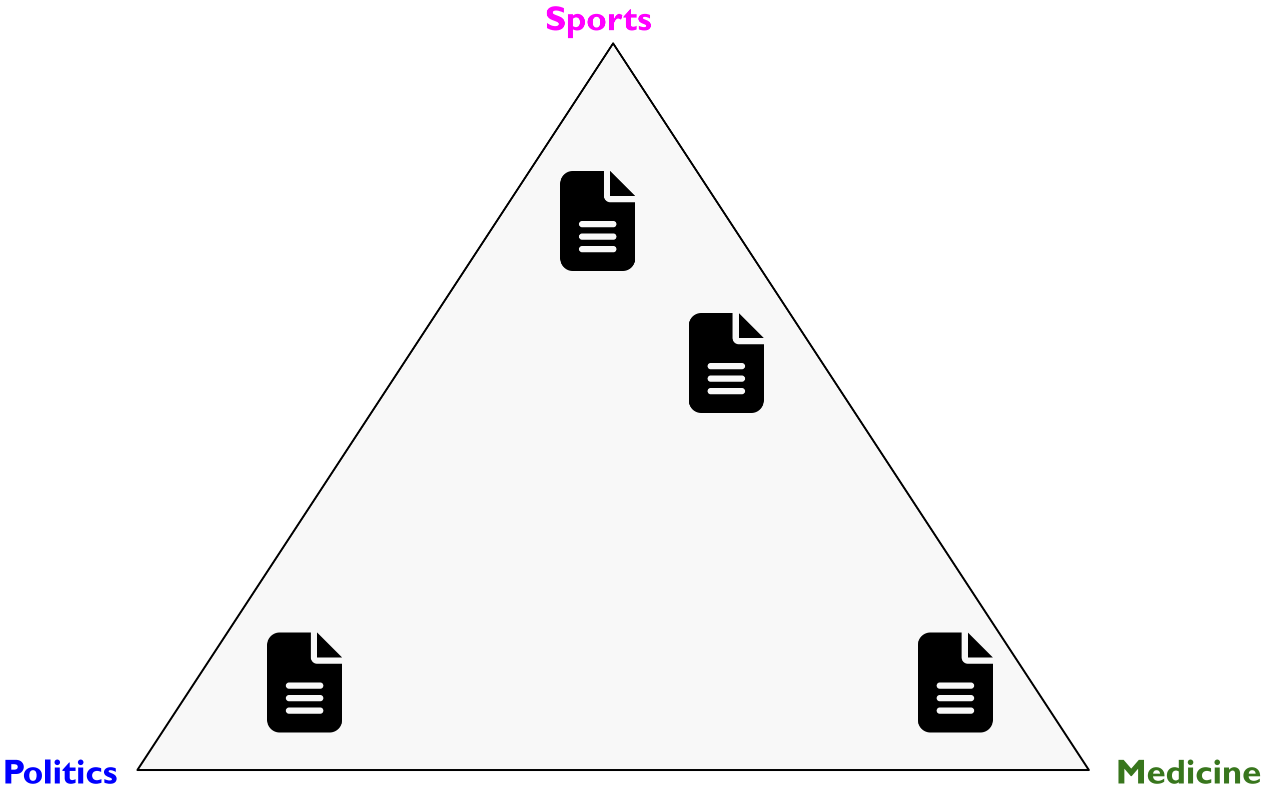
LDA’s approach:

- Each **document** is a collection of **topics** in a certain proportion
- Each **topic** is a collection of **words** in a certain proportion

CDHU

UPPSALA
UNIVERSITET

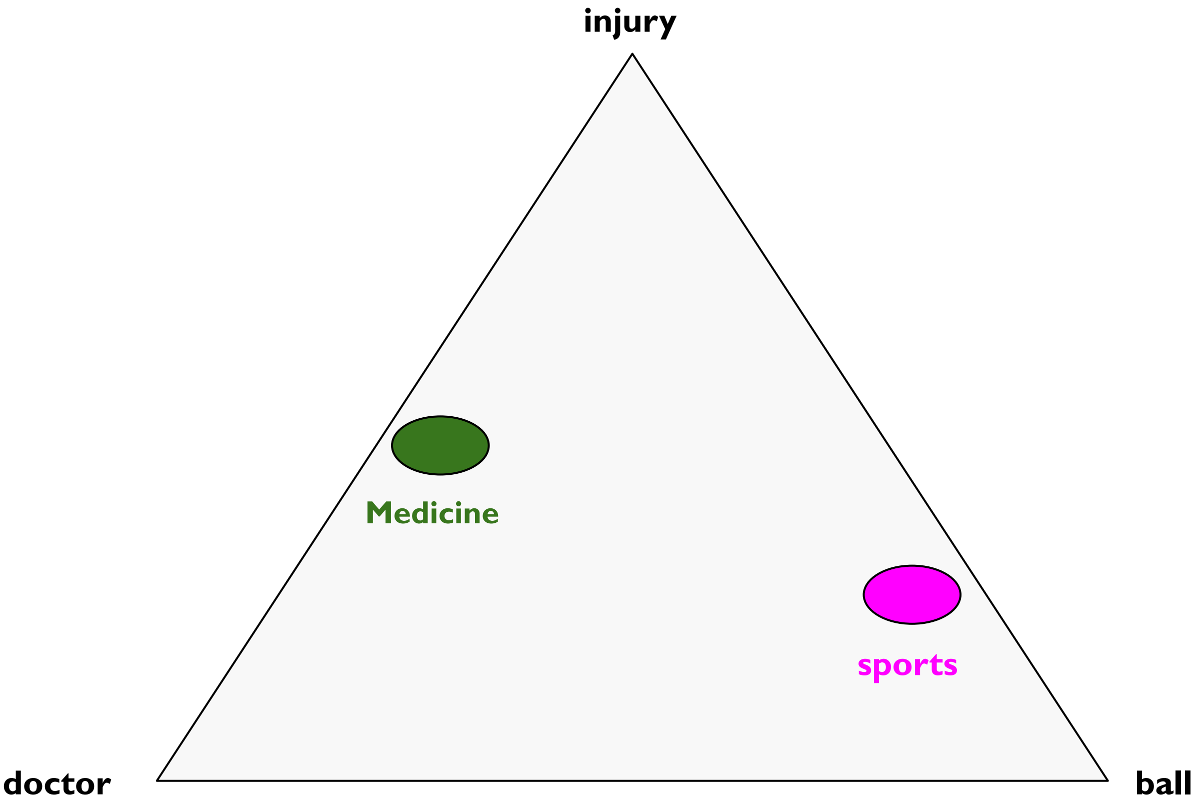
“Each **document** is a collection of **topics** in a certain proportion”



CDHU

UPPSALA UNIVERSITET

“Each **topic** is a collection of **words** in a certain proportion”



What is Topic Modelling

- Approach to discover hidden semantic patterns in a text corpus
- unsupervised machine learning to analyze and identify clusters or groups of similar words within a body of text
- Topic modelling is a type of statistical model used for discovering abstract topics within a collection of documents. These models can help in summarizing large datasets of textual information by categorizing documents into topics.
-
- A generative statistical model to find hidden relations between documents, words, topics (groups of words)
- where we have a large collection of text but don't really know the nature of its contents, topic models can help us get a glimpse inside and identify the main themes in our corpus
- it's important to remember that these algorithms cannot guarantee that the words in each topic will be related to one another conceptually — only that they frequently occur together in your data for some reason.
- topic modeling algorithms are great at identifying clusters of words that frequently co-occur, they do not actually understand the context in which those words occur.

CDHU

UPPSALA UNIVERSITET

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

doctor
injury
injury
injury
doctor

Guess the topics!

CDHU

UPPSALA UNIVERSITET

Doc 1

politician
election
election
ball
politician

Politics

Doc 2

ball
injury
injury
ball
ball

Sports

Doc 3

politician
election
politician
election
ball

Politics

Doc 4

doctor
injury
injury
injury
ball

Medicine

CDHU

UPPSALA
UNIVERSITET

Doc 1

ioadkfjlk
abadlihe
abadlihe
ylsk
ioadkfjlk

Doc 2

ylsk
ölskjdlfä
ölskjdlfä
ylsk
ylsk

Doc 3

ioadkfjlk
abadlihe
ioadkfjlk
abadlihe
ylsk

Doc 4

llleijad
ölskjdlfä
ölskjdlfä
ölskjdlfä
llleijad

Guessing the topics now is hard…

How can we solve this problem?

CDHU

UPPSALA
UNIVERSITET

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

doctor
injury
injury
injury
doctor

Topic 1 **Topic 2** **Topic 3**

Let's find 3 topics in the documents

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

doctor
injury
injury
injury
doctor

Topic 1 **Topic 2** **Topic 3**

Label/color every word with a topic: "Bottom-up approach"

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

doctor
injury
injury
injury
doctor

Topic 1 **Topic 2** **Topic 3**

Label/color every word in the documents with a topic

Doc 1

politician
election
election
ball
politician

80 % Topic 1
20 % Topic 2

Doc 2

ball
injury
injury
ball
ball

80 % Topic 2
20 % Topic 3

Doc 3

politician
election
politician
election
ball

60 % Topic 1
20 % Topic 2
20 % Topic 3

Doc 4

doctor
injury
injury
injury
doctor

80 % Topic 3
20 % Topic 2

Topic 1 **Topic 2** **Topic 3**

CDHU

UPPSALA
UNIVERSITET

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

doctor
injury
injury
injury
doctor

How does a good coloring look like?

Doc 1

politician
election
election
ball
politician

How does a good coloring look like?

1) Coloring of each **document** should be as homogenous as possible

CDHU

UPPSALA UNIVERSITET

politician	ball	election	injury	doctor
politician	ball	election	injury	doctor
politician	ball	election	injury	
politician	ball	election	injury	
	ball		injury	

How does a good coloring look like?

1) Coloring of each **document** should be as homogenous as possible
2) Coloring of each **word** should be as homogeneous as possible

CDHU

UPPSALA UNIVERSITET

LDA Algorithm Intuition

CDHU

UPPSALA
UNIVERSITET

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

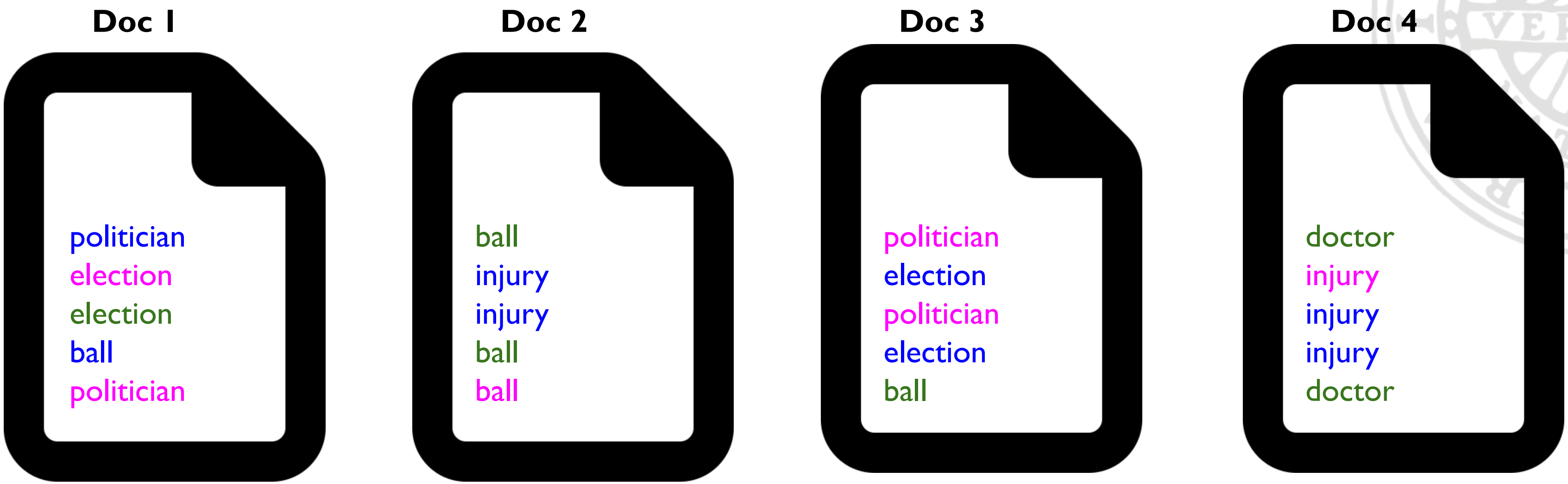
doctor
injury
injury
injury
doctor

Topic 1 **Topic 2** **Topic 3**

1) Select number of topics (hyperparameter)

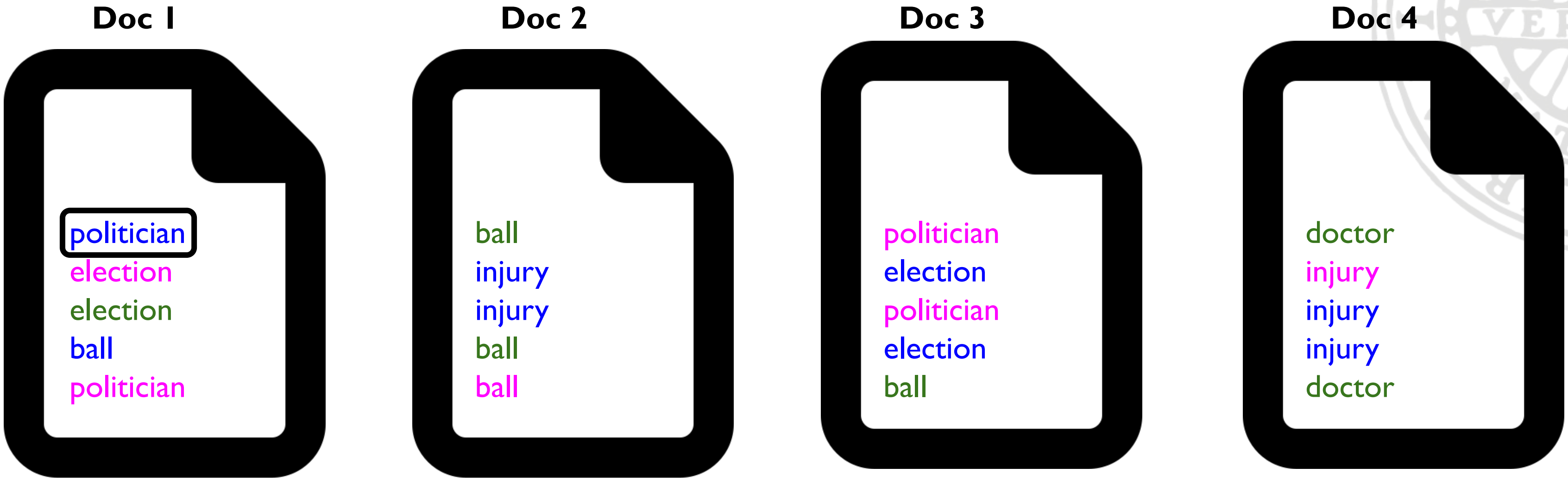
CDHU

UPPSALA UNIVERSITET



Topic 1 **Topic 2** **Topic 3**

1) Select number of topics (hyperparameter)
2) Start with random coloring of words in documents

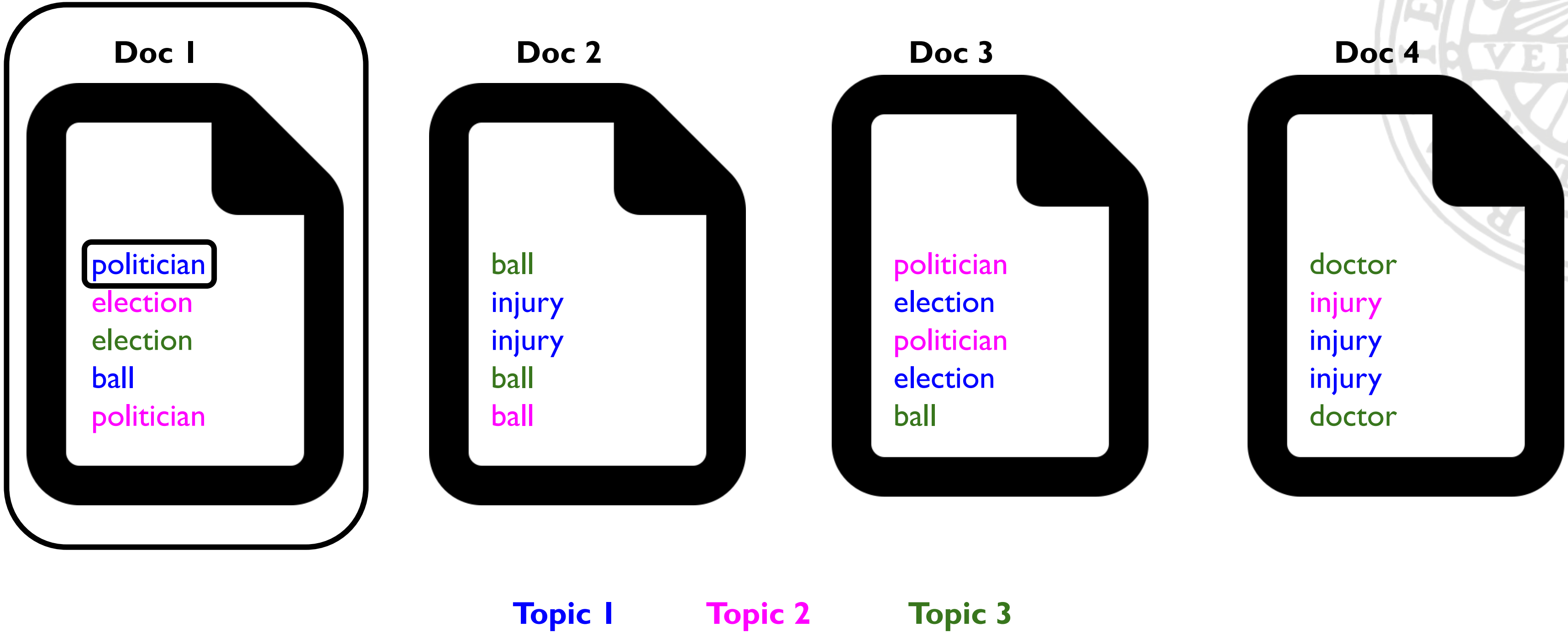


Topic 1 **Topic 2** **Topic 3**

1) Select number of topics (hyperparameter)
2) Start with random coloring/topic for words in documents
3) Iterate through every word and update coloring/topic

CDHU

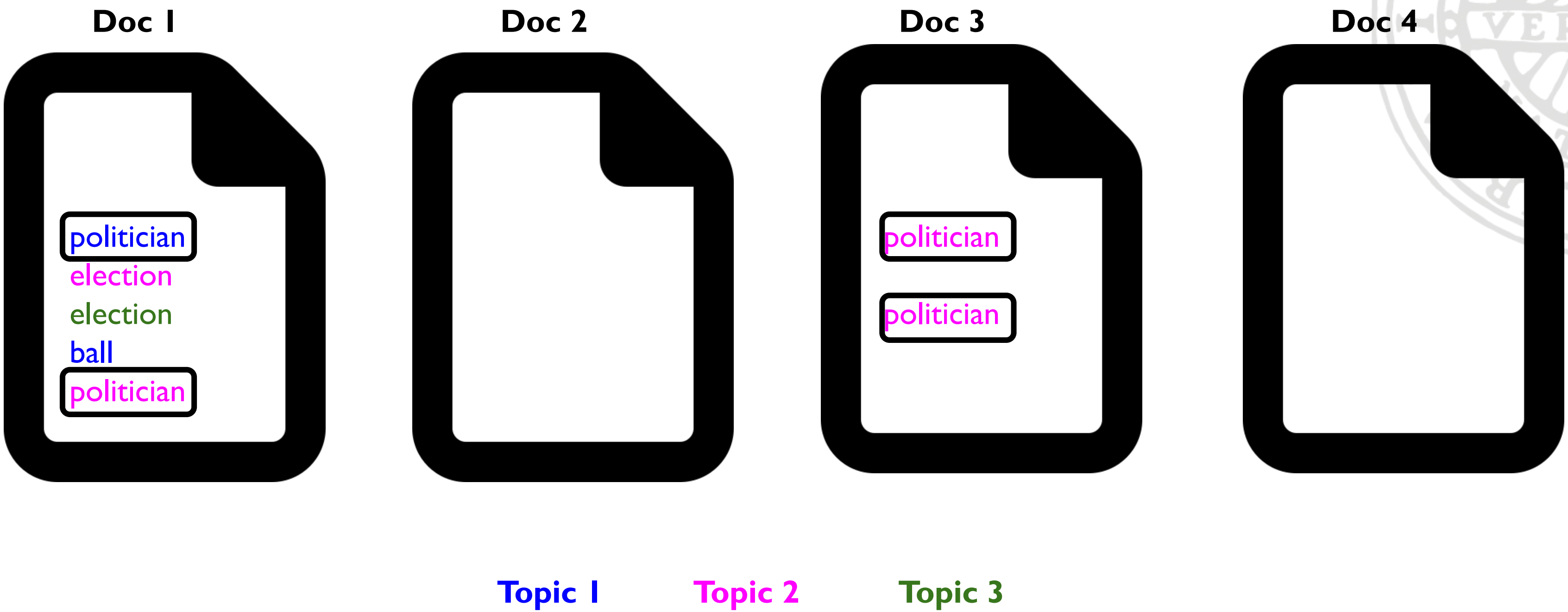
UPPSALA UNIVERSITET



1) Look at coloring/topics of words in same document

CDHU

UPPSALA UNIVERSITET

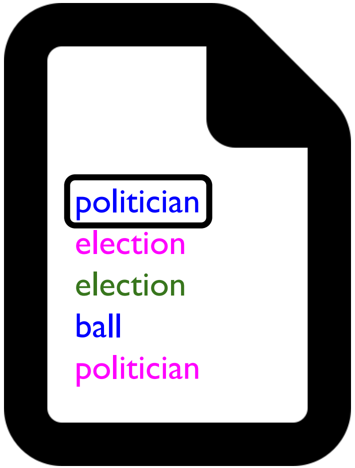


1) Look at coloring/topics of words in same document
2) Look at coloring/topics of same words in all documents

CDHU

UPPSALA UNIVERSITET

Doc 1



politician
politician
politician
politician

1) Look at topics of words in same document

40% Topic 2
40% Topic 1
20% Topic 3

2) Look at topics of same words in all documents

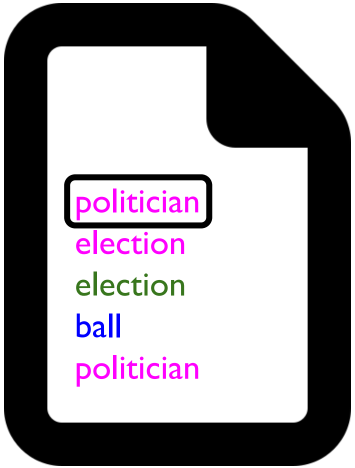
75% Topic 2
25% Topic 1

Most assigned topic/color in same document and for all same words "politician": Topic 2
→ assign Topic 2 to the current word "politician"

CDHU

UPPSALA UNIVERSITET

Doc 1



politician
politician
politician
politician

1) Look at topics of words in same document

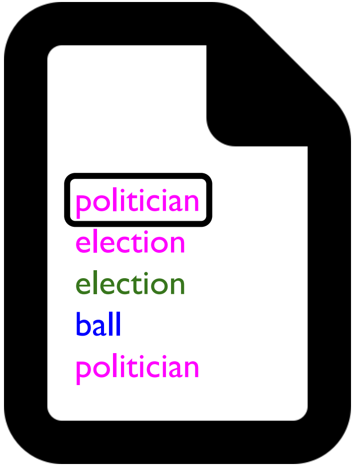
40% Topic 2
40% Topic 1
20% Topic 3

2) Look at topics of same words in all documents

75% Topic 2
25% Topic 1

Most assigned topic/color in same document and for all same words "politician": Topic 2
→ assign Topic 2 to the current word "politician"

Doc 1



politician
politician
politician

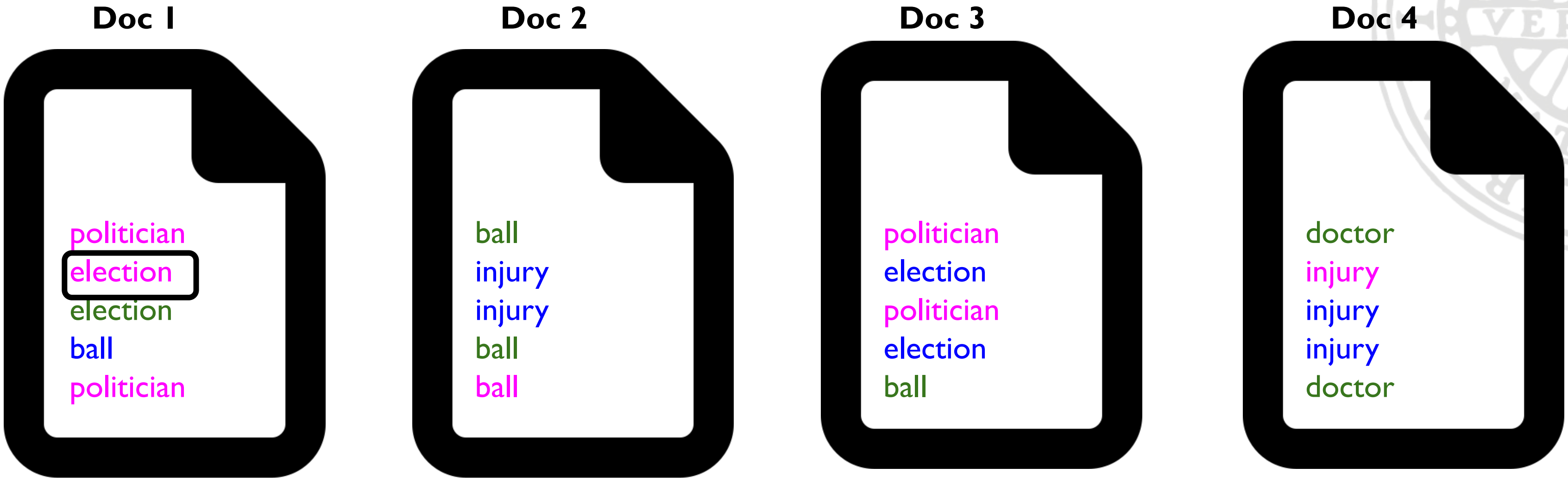
1) Look at topics of words in same document

40% Topic 1
40% Topic 2
20% Topic 3

2) Look at topics of same words in all documents

75% Topic 2
25% Topic 1

Most assigned topic/color in same document and for all same words “politician”: Topic 2
→ assign Topic 2 to the current word “politician”



Topic 1 **Topic 2** **Topic 3**

1) Select number of topics (hyperparameter)
2) Start with random coloring of words in documents
3) Iterate through every word and update coloring

After we repeat this process for a few iterations

….

CDHU

UPPSALA
UNIVERSITET

Doc 1	Doc 2	Doc 3	Doc 4
politician	ball	politician	doctor
election	injury	election	injury
election	injury	politician	injury
ball	ball	election	injury
politician	ball	ball	doctor

Topic 1
politician (4)
election (4)

0.5* "politician"
+ 0.5* "election"

Topic 2
ball (5)
injury(2)

0.71* "ball"
+ 0.29* "injury"

Topic 3
injury (3)
doctor (2)

0.6*"injury"
+ 0.4*"doctor"

Doc 1

politician
election
election
ball
politician

Doc 2

ball
injury
injury
ball
ball

Doc 3

politician
election
politician
election
ball

Doc 4

doctor
injury
injury
injury
doctor

Topic 1: Politics
politician (4)
election (4)

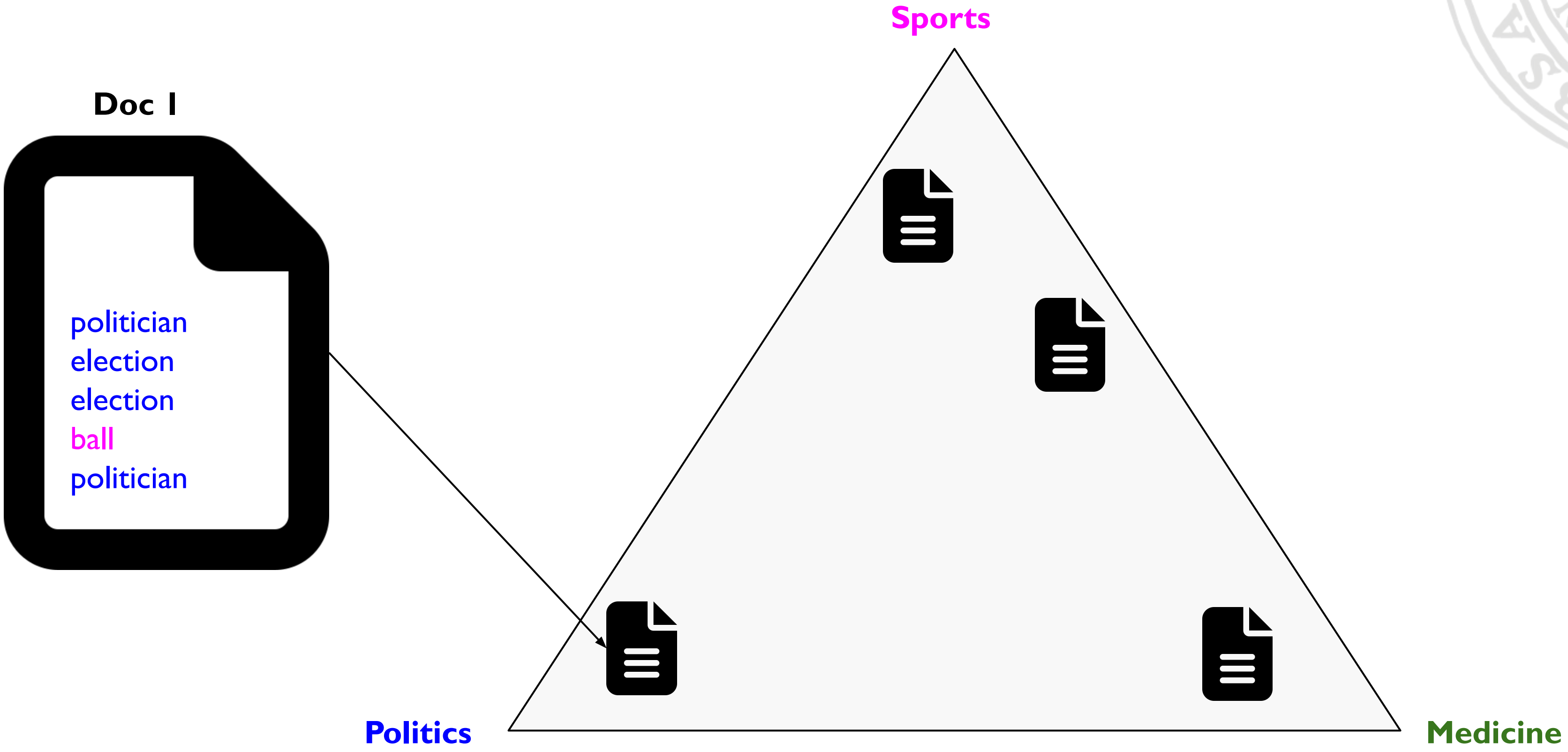
Topic 2: Sports
ball (5)
injury(2)

Topic 3: Medicine
injury (3)
doctor (2)

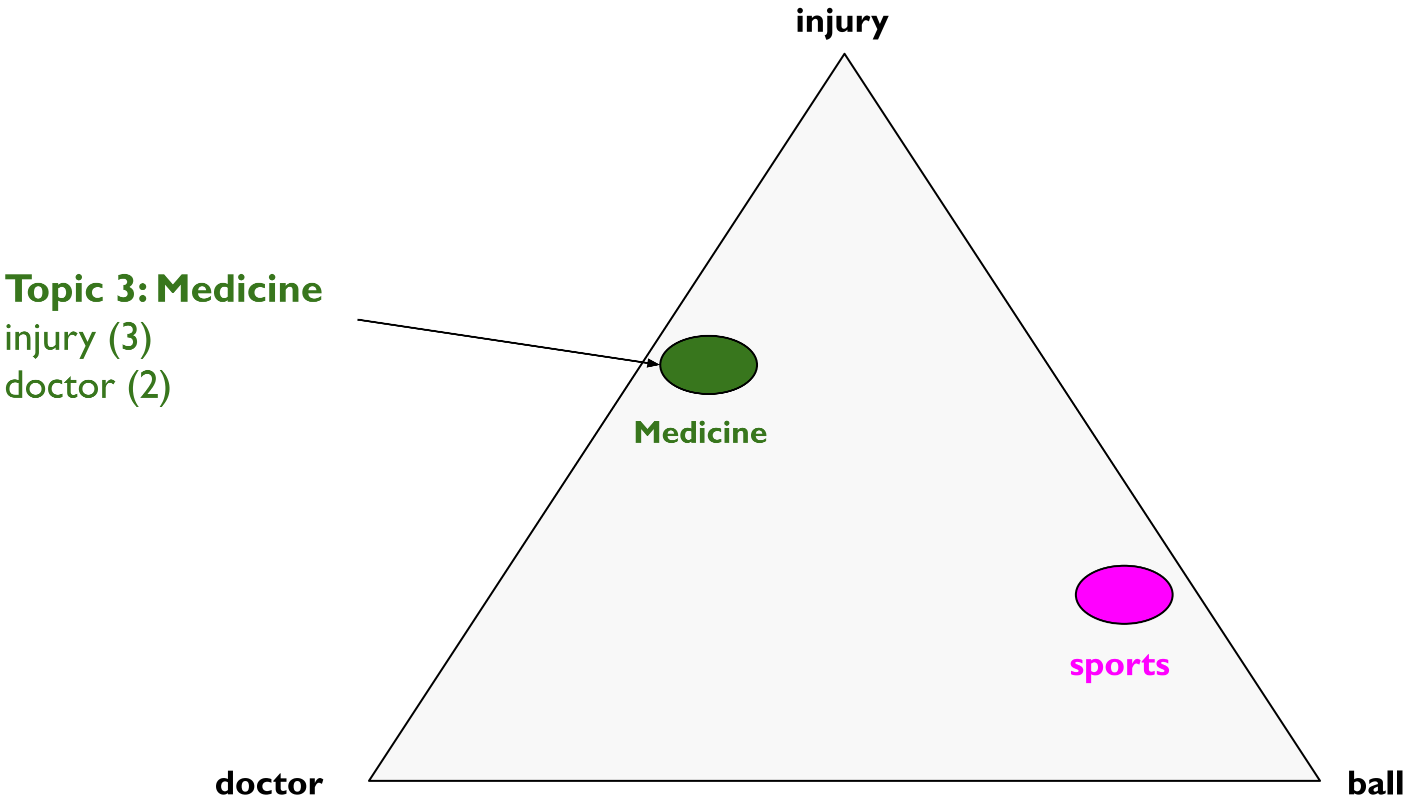
CDHU

UPPSALA UNIVERSITET

Documents - Topics Dirichlet Distribution



Topics - Words Dirichlet Distribution



CDHU

UPPSALA UNIVERSITET

Latent Dirichlet Allocation (LDA)

Notes:

- Topics are collections of words, Documents are collections of topics
- The same topic can be assigned to multiple documents
- The same words can be part of multiple topics
- LDA is an unsupervised learning algorithm
    - We only need to select the number of topics we want to find ("hyperparameter")
- The algorithm is a bit complicated to implement. We will use the Python's Gensim library to use topic modeling
- We want to apply topic modeling to more complicated texts

CDHU

UPPSALA UNIVERSITET

Digital humanities (**DH**) is an area of scholarly activity at the intersection of computing or digital technologies and the disciplines of the humanities. It includes the systematic use of digital resources in the humanities, as well as the analysis of their application.[1][2] DH can be defined as new ways of doing scholarship that

- We don't want to consider brackets "(" or other special characters like "[1]"
- "Digital" and "digital" should be the same word
- What about "includes", "including", "include"?
- Are words like "a", "and", or "of" relevant for a topic?

CDHU

UPPSALA UNIVERSITET

Text Preprocessing Steps

1. Remove special characters (!=?:"+)
2. Change all words to lowercase
3. Remove stop words ("and", "a", "or")
4. Remove single letters ("R", "t")
5. Tokenize text (split text into tokens)
   - "this is an example" → ["this", "is", "an", "example"]
6. Lemmatization
   - "goes", "go", "going" → "go"

CDHU

UPPSALA UNIVERSITET

Questions before the Demo?

CDHU

UPPSALA
UNIVERSITET

Demo

I want to thank Luis Serrano (https://serrano.academy/) and his incredible explanation of topic modeling that inspired this presentation

CDHU
UPPSALA UNIVERSITET