Natural Language Processing for Trend Forecasting

Heládio Sampaio Lopes

Computer Engineering (ITA 2020)

Advisor: Filipe Alves Neto Verri

CONTENT

CONTENT

INTRODUCTION

Natural Language Processing can perform a lot of tasks, such as identifying different topics for a set of documents, classifying texts on predefined subjects, and beyond that extract the sentiment to know what people are saying about something.

Motivation

Over the years, **more and more knowledge** is **generated** and we humans are not able to process such an amount of information. **Natural Language Processing** emerges as a technology capable of **assisting** us **in this** hard **task**.

Objectives

Explore Natural Language Processing techniques to propose a framework to **modeling in real-time the topics'** evolution, and evaluate these models ability in **modeling** the themes occurrence **over the years**.

INTRODUCTION | Related Works

Finding meaningful topics in a document collection has been used for a lot of authors for the most various applications like forecast trends. Predicting future trends can be very helpful in various applications, like to model the evolution of research.

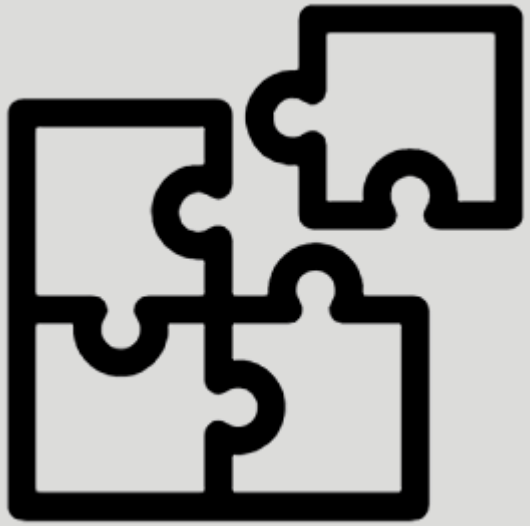
Topic Modeling

Hurtado *et al.* (2016) use topic modeling to inspect research publications to model the evolution of the direction of research and forecast the near future trends in IT industry.

Jelodar *et al.* (2020) recently use topic modeling in Reddit related posts about the new disease Covid-19 to group similar comments and perform a sentiment analysis.

INTRODUCTION | Gap

Topic discovery and trend forecast were subjects widely explored in the literature. With that in mind, we wish, in this work, to reproduce these techniques. However, in addition to what has been presented we want to be able to explore some modifications.



Real-time system that will keep receiving news in a continuous process.

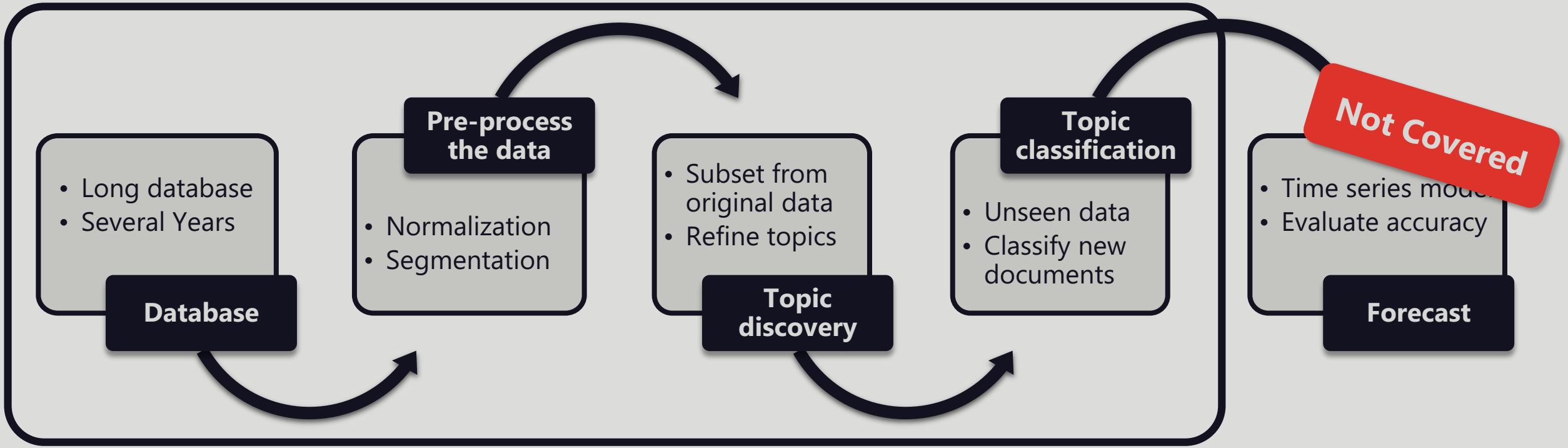
Redo the discovery process will demand an expensive computational cost.

New topic classification-based system will be proposed in order fill this gap.

CONTENT

METHODOLOGY AND RESULTS

We want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.



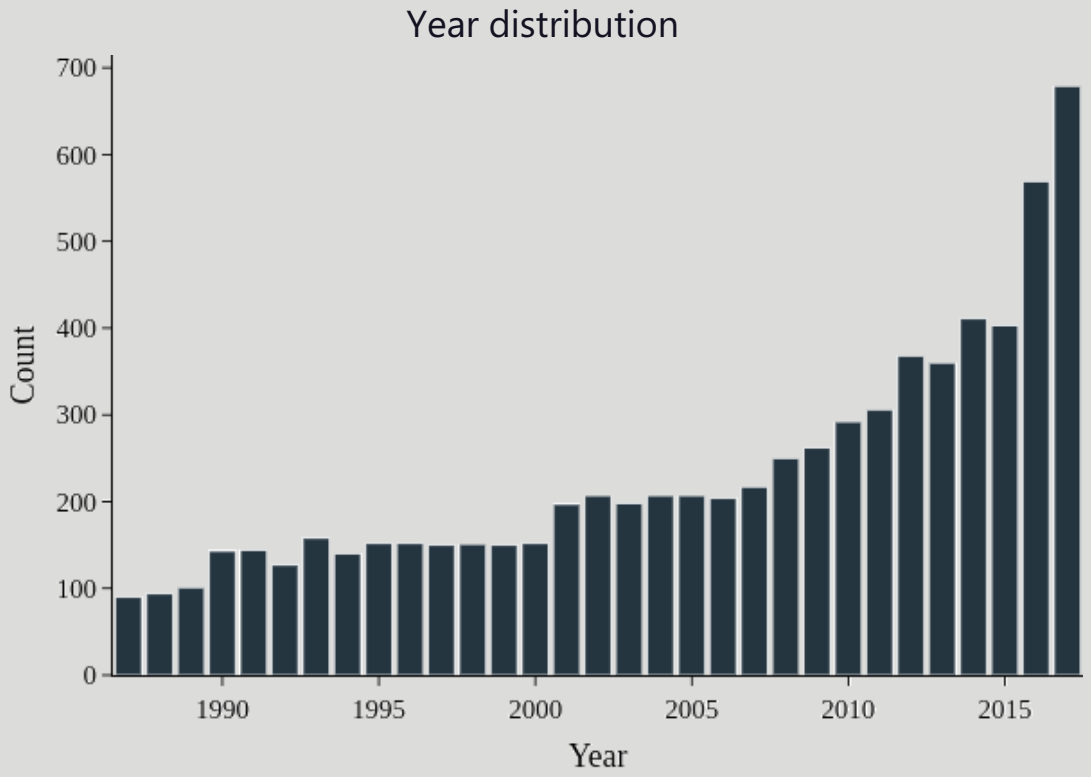
CONTENT

METHODOLOGY AND RESULTS |

With over than seven thousand entries spread in a thirty-year range, we use the database of Neural Information Processing System, one of the most prestigious yearly events in machine learning community.

NEURAL INFORMATION PROCESSING SYSTEMS

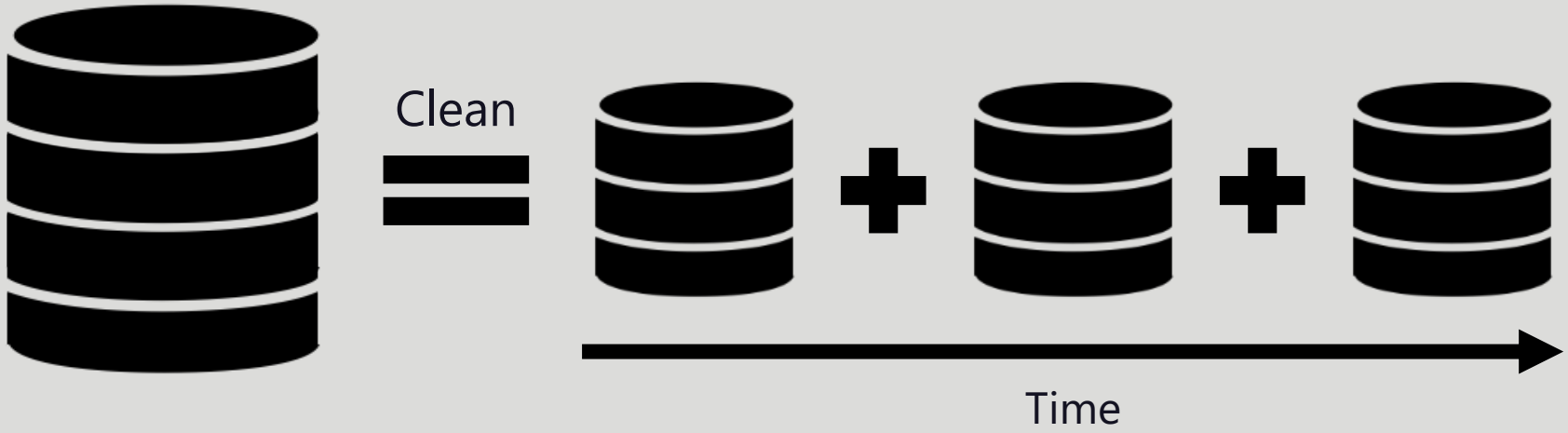
- Academic Papers
- 7241 Documents
- Yearly Conference (1987 – 2017)



CONTENT

METHODOLOGY AND RESULTS |

We must define a data pre-processing pipeline to normalize the documents, putting all of them at the same pattern. After processing the data, we need to index them over time and, then, split the full treated data set into three subsets.



METHODOLOGY AND RESULTS |

Pre-processing Steps

- Drop links
- Remove numbers
- Expand contractions
- Remove punctuation
- Convert special characters
- Case Conversion
- Lemmatization
- Remove stop-words

Raw Text

767\n\nSELF-ORGANIZATION OF ASSOCIATIVE DATABASE\nAND ITS APPLICATIONS\nHisashi Suzuki and Suguru Arimoto\nOsaka University, Toyonaka, Osaka 560, Japan\nABSTRACT\nAn efficient method of self-organizing associative databases is proposed together with\napplications to robot eyesight systems. The proposed databases can associate any input\nwith some output. In the first half part of discussion, an algorithm of self-organization is\nproposed. From an aspect of hardware, it produces a new style of neural network. In the\nlatter half part, an applicability to handwritten letter recognition and that to an autonomous\nmobile robot system are demonstrated.\n\n

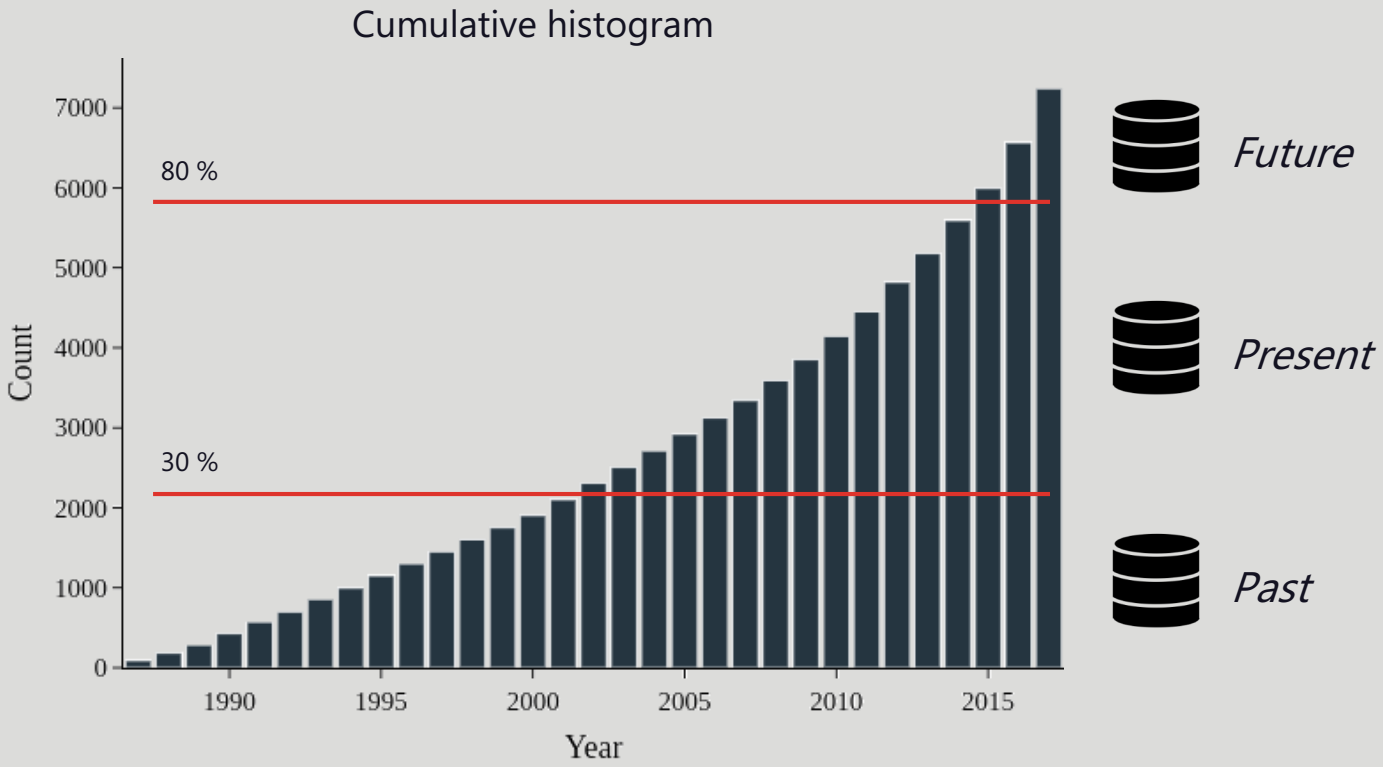
Normalization

Processed

self organization associative database application hisashi suzuki suguru arimoto osaka university toyonaka osaka japan abstract efficient method self organize associative database propose together application robot eyesight system propose database associate input output first half part discussion algorithm self organization propose aspect hardware produce style neural network latter half part applicability handwritten letter recognition autonomous mobile robot system demonstrate

Demonstration about the normalization process

METHODOLOGY AND RESULTS |

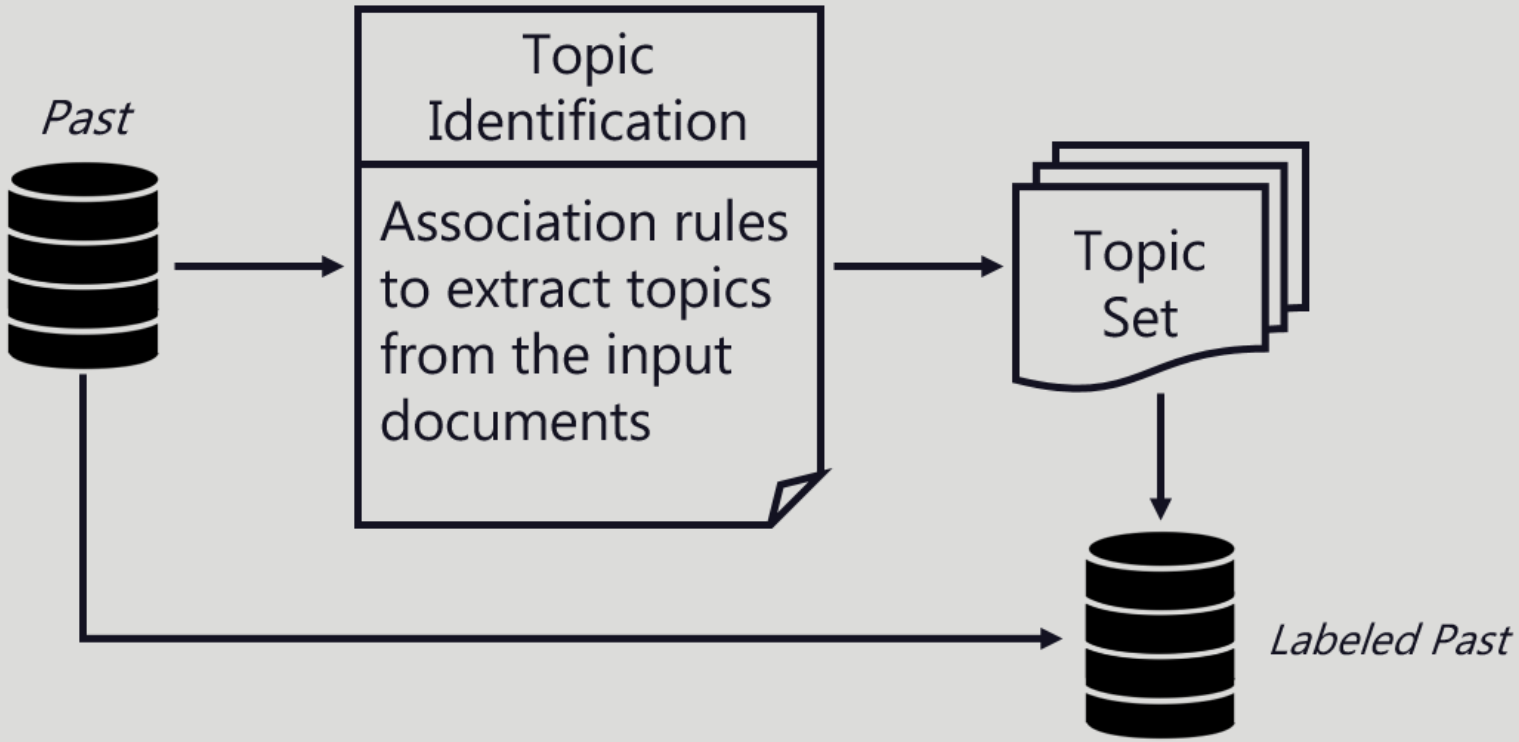


Subset	Nº of documents	Years
Future	1248	2016 – 2017
Present	3685	2003 – 2015
Past	2308	1987 – 2002

CONTENT

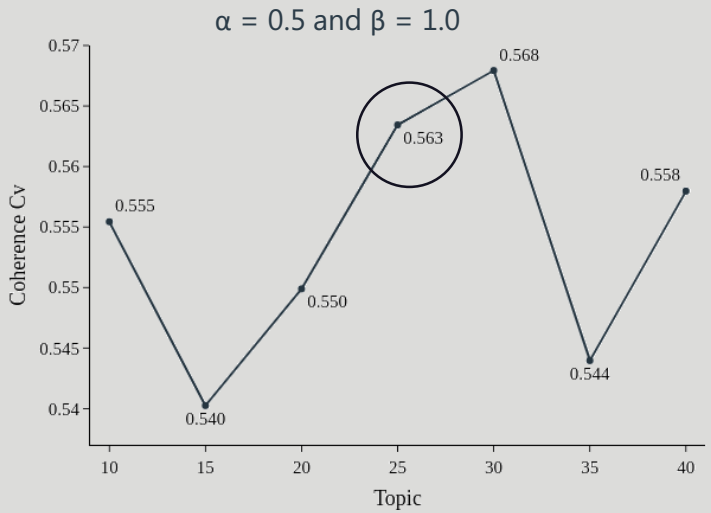
METHODOLOGY AND RESULTS |

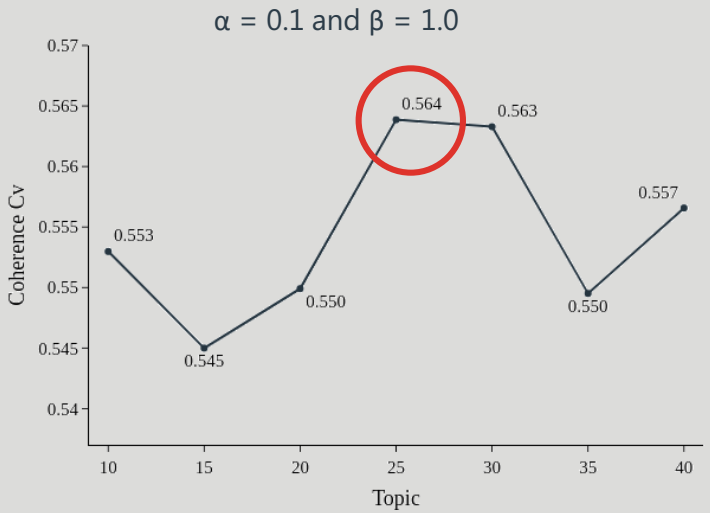
With the *Past* set, a topic modeling technique will be applied to find the discussed subjects in the documents. Next to the topic identification, we will obtain a topic set, then each document contained in *Past* can be labeled with at least one topic.

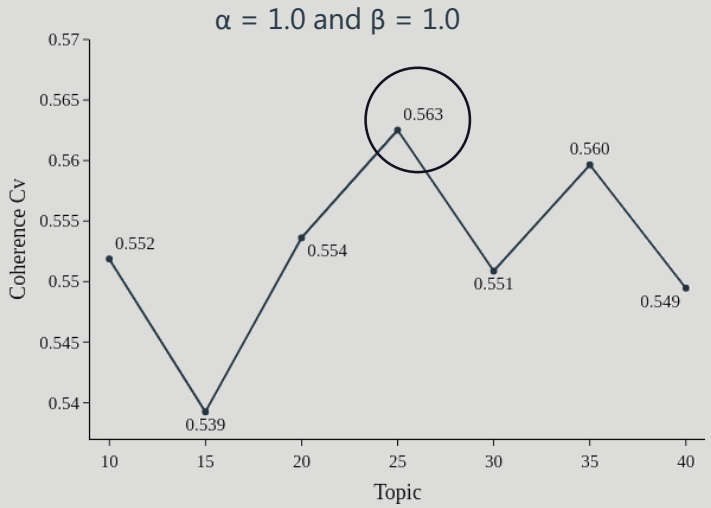


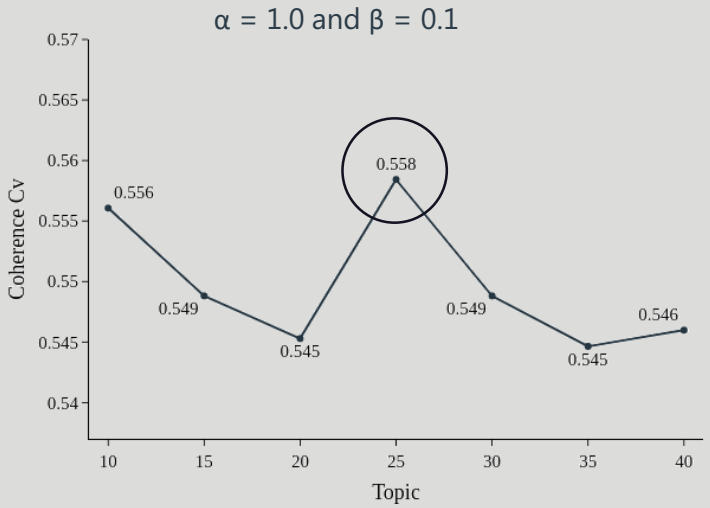
Topic Coherence Optimization

- Find the best topic models
- Maximize coherence score C_v
- Grid search on α, β and K
  - $\alpha \in \{0.1, 0.5, 1\}$
  - $\beta \in \{0.05, 0.1, 0.5, 1\}$
  - $K \in \{10, 15, 20, 25, 30, 35, 40\}$

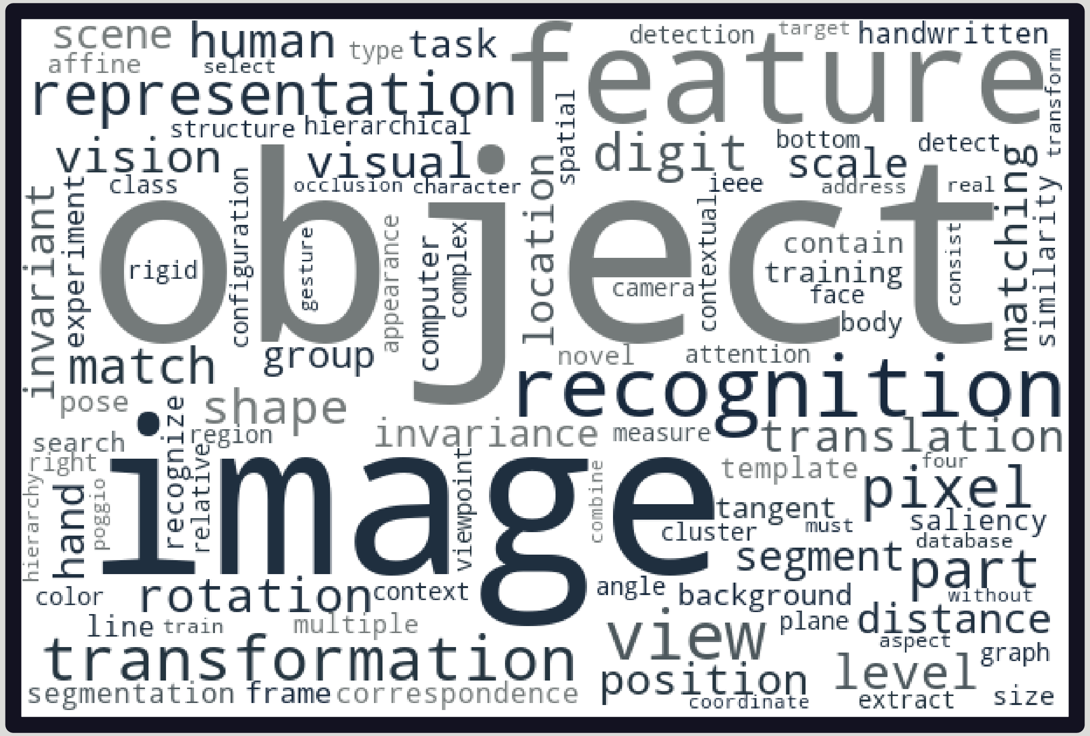


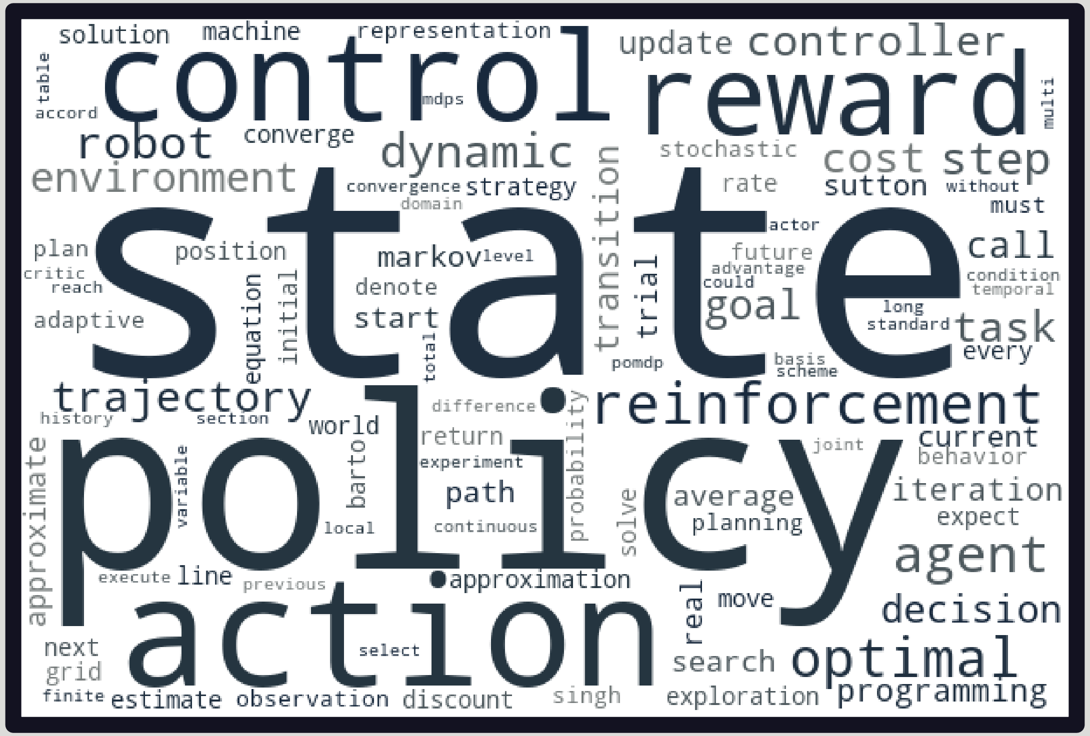






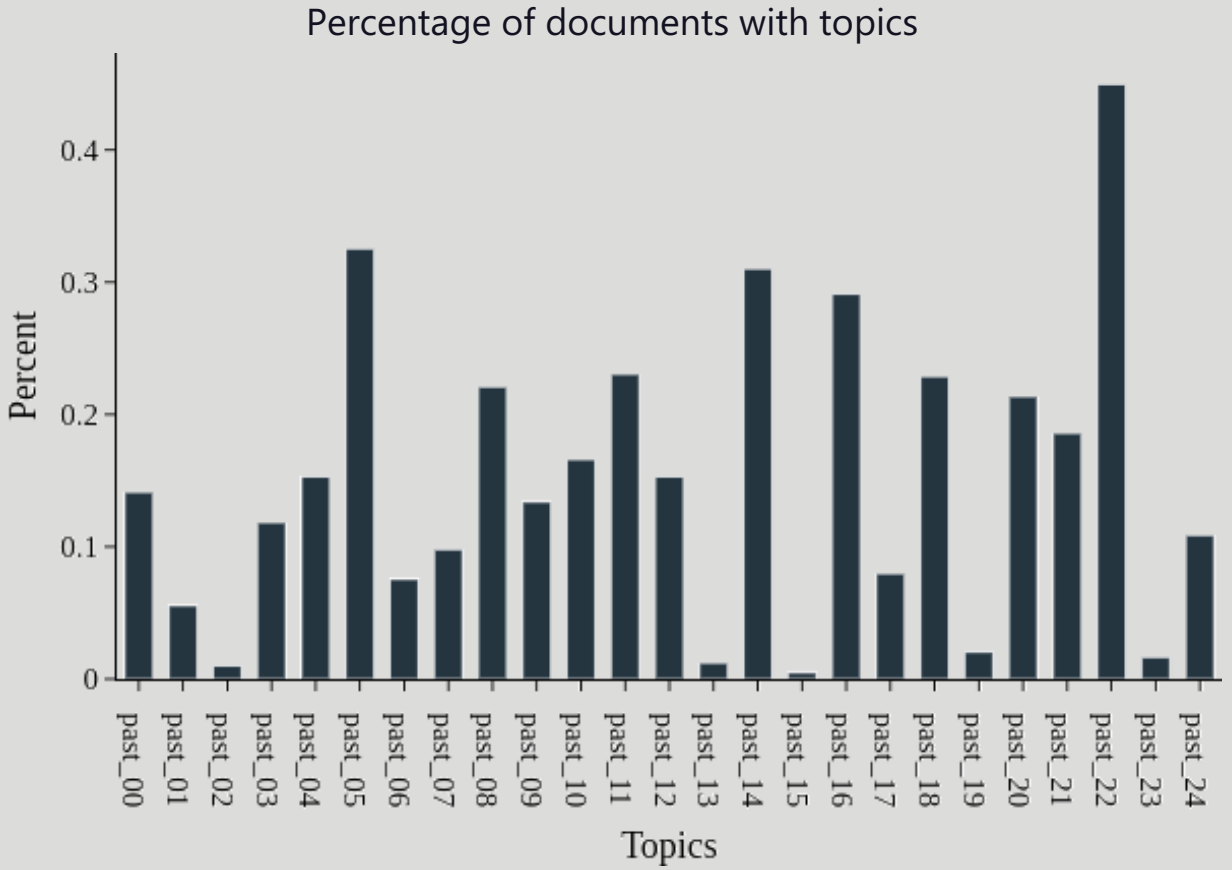
Past Topics


object
image
feature
recognition
representation
transformation
vision
visual
digit
scale
match
shape
view
rotation
translation
pixel
part
distance
level
position
segment
invariance
invariant
location
matching
similarity
scene
human
task
type
detection
target
handwritten
affine
select
structure
hierarchical
bottom
detect
spatial
transform
class
occlusion
character
ieee
address
real
experiment
configuration
rigid
gesture
appearance
computer
complex
camera
contextual
consist
contain
training
face
body
group
novel
attention
pose
search
region
measure
four
template
right
recognize
relative
hierarchy
hand
poggio
viewpoint
combine
tangent
cluster
must
saliency
database
color
context
angle
background
without
line
train
multiple
plane
aspect
graph
segmentation
frame
correspondence
coordinate
extract
size


state
policy
action
control
reward
controller
update
solution
machine
representation
table
accord
mdps
multi
robot
converge
dynamic
stochastic
cost
step
environment
convergence
strategy
domain
rate
sutton
without
must
transition
actor
plan
position
markov
level
future
advantage
could
call
condition
temporal
critic
reach
denote
goal
long
standard
task
adaptive
equation
initial
start
trial
total
pomdp
basis
scheme
every
trajectory
history
section
world
difference
reinforcement
return
experiment
probability
joint
current
behavior
iteration
expect
approximate
variable
barto
path
average
planning
local
continuous
solve
agent
execute
line
previous
approximation
real
move
decision
next
grid
select
search
optimal
finite
estimate
observation
discount
singh
exploration
programming

Documents *vs.* Topics

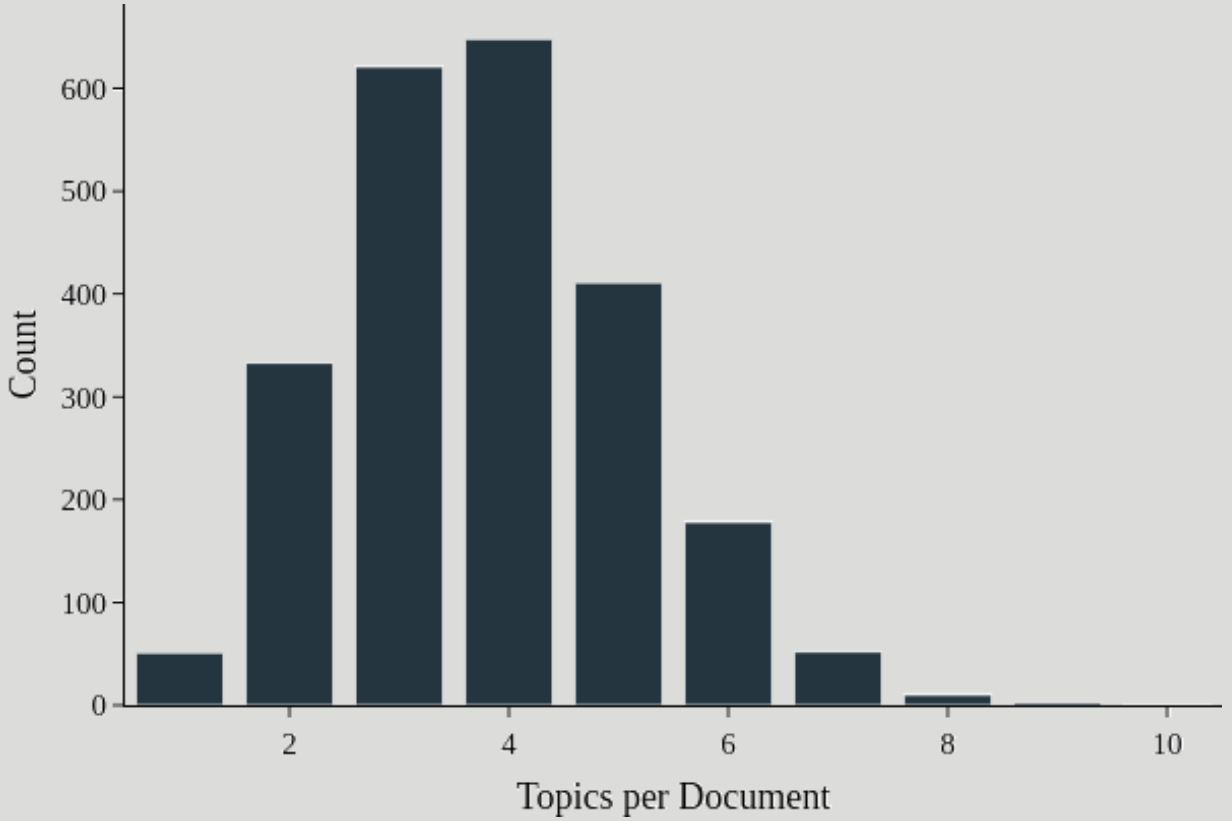
- Topic distribution per document
- At least one topic per document
- Document can cover more than one topic
- Threshold for topic in document
  - $^{1}/_{K} = 8\ \%$



Documents *vs.* Topics

- Topic distribution per document
- At least one topic per document
- Document can cover more than one topic
- Threshold for topic in document
  - $^{1}/_{K} = 8\ \%$

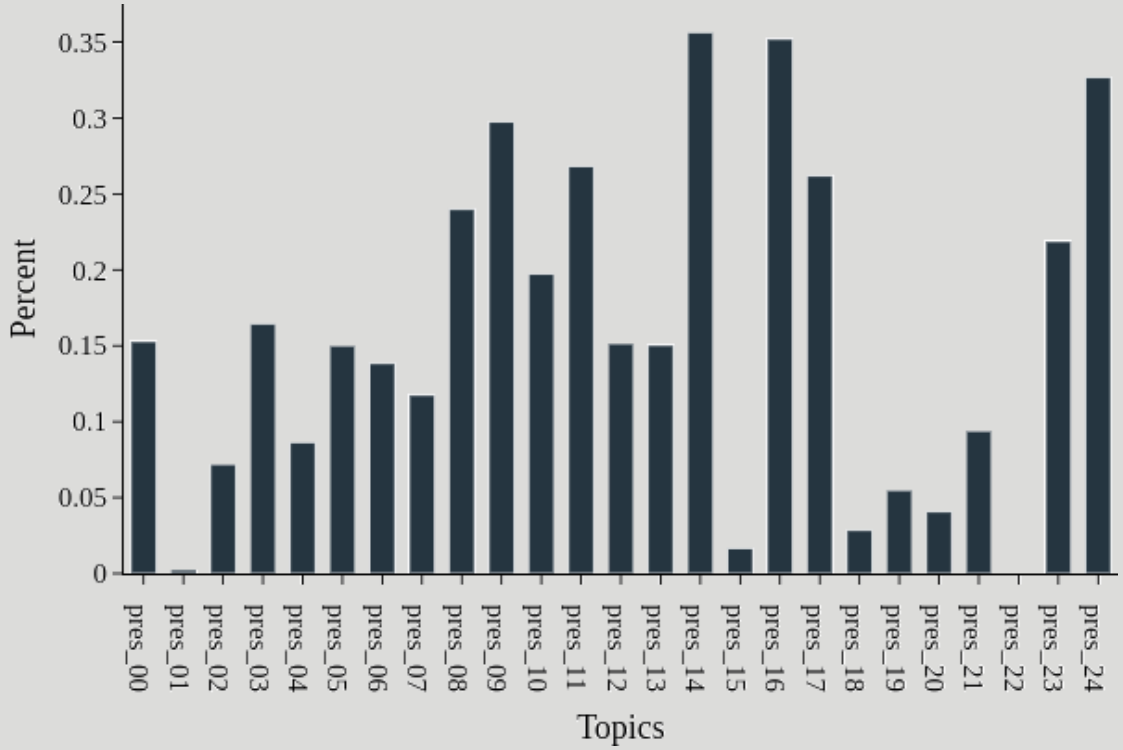
Histogram for number of topics per documents



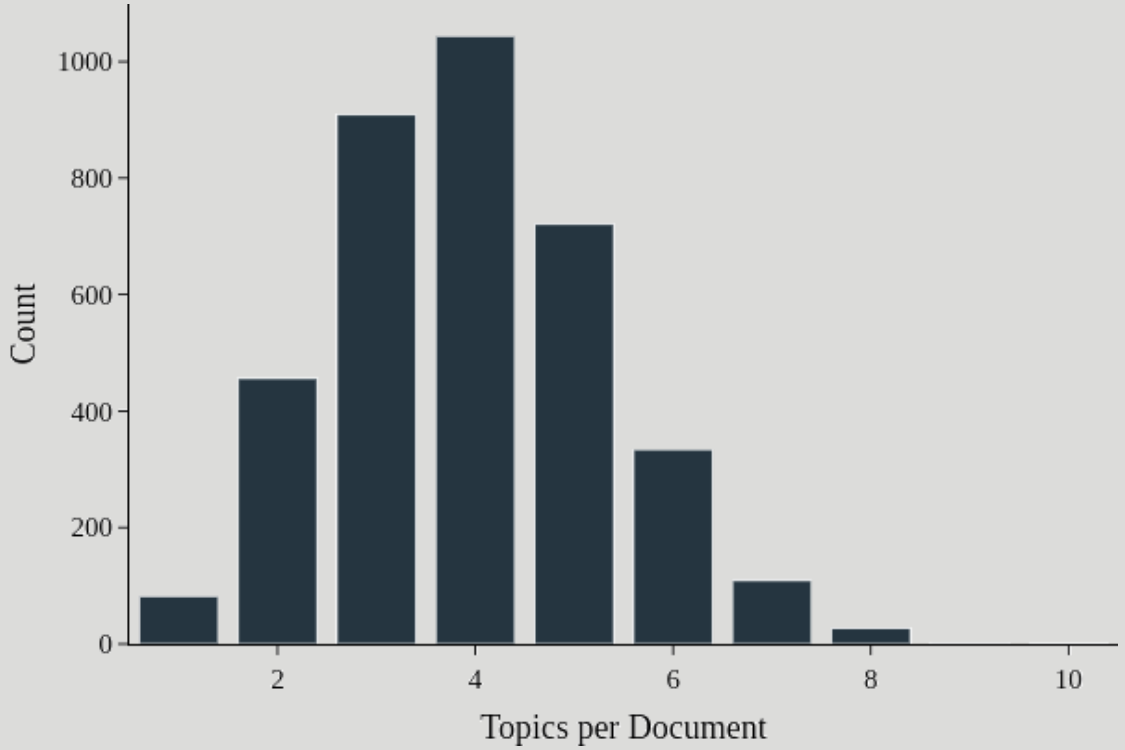
Present Topics

METHODOLOGY AND RESULTS |

Percentage of documents with topics



Histogram for number of topics per documents



Past-Present Combination

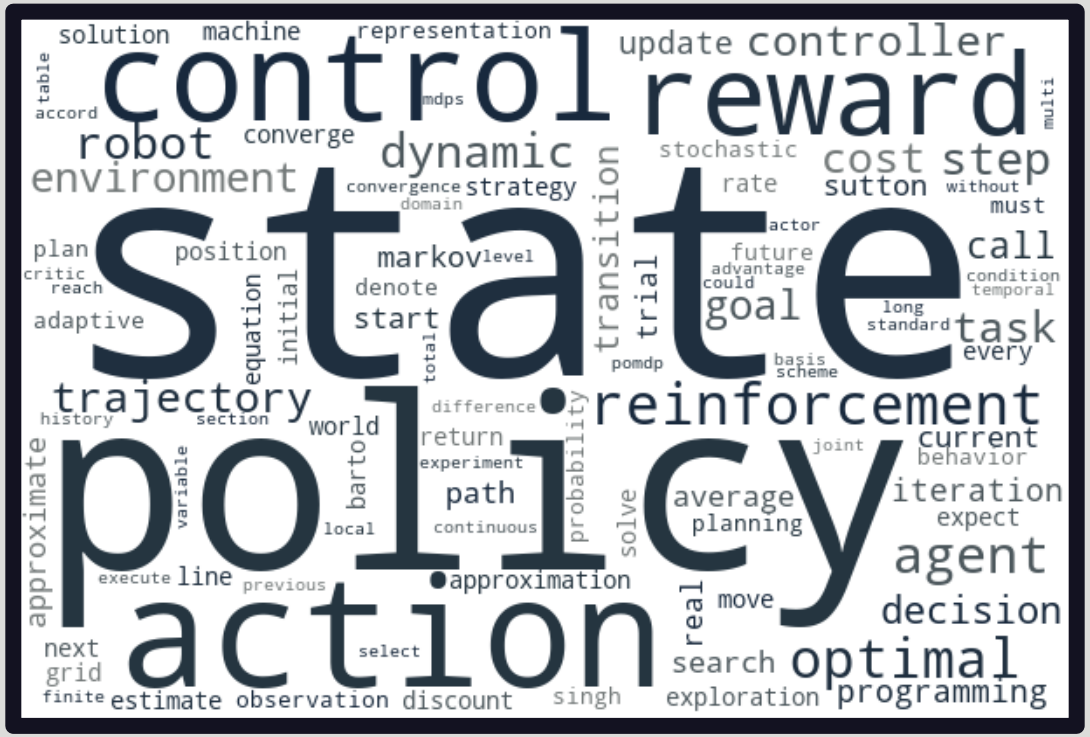
- Evaluate past models over present
- Combination metrics
  - $Sim_1 = \frac{|Top_N(Past_i) \cap Top_N(Pres_j)|}{N}$
  - $Sim_2 = \sqrt{\sum \left(w_v^{Past_i} - w_v^{Pres_j} \right)}$
  - Sim_1 / Sim_2

Past	Present	Sim_1	Sim_2	Sim_1 / Sim_2
11	06	58 %	0.026	1119.71
10	05	56 %	0.031	911.39
14	16	44 %	0.040	555.48
20	09	38 %	0.036	528.22
18	11	40 %	0.043	467.59
07	12	34 %	0.038	445.64
04	02	54 %	0.061	440.94
00	08	46 %	0.055	417.94
05	03	32 %	0.047	341.65
22	07	36 %	0.057	343.72

Past Topic 11

Present Topic 06

Past Topic 10

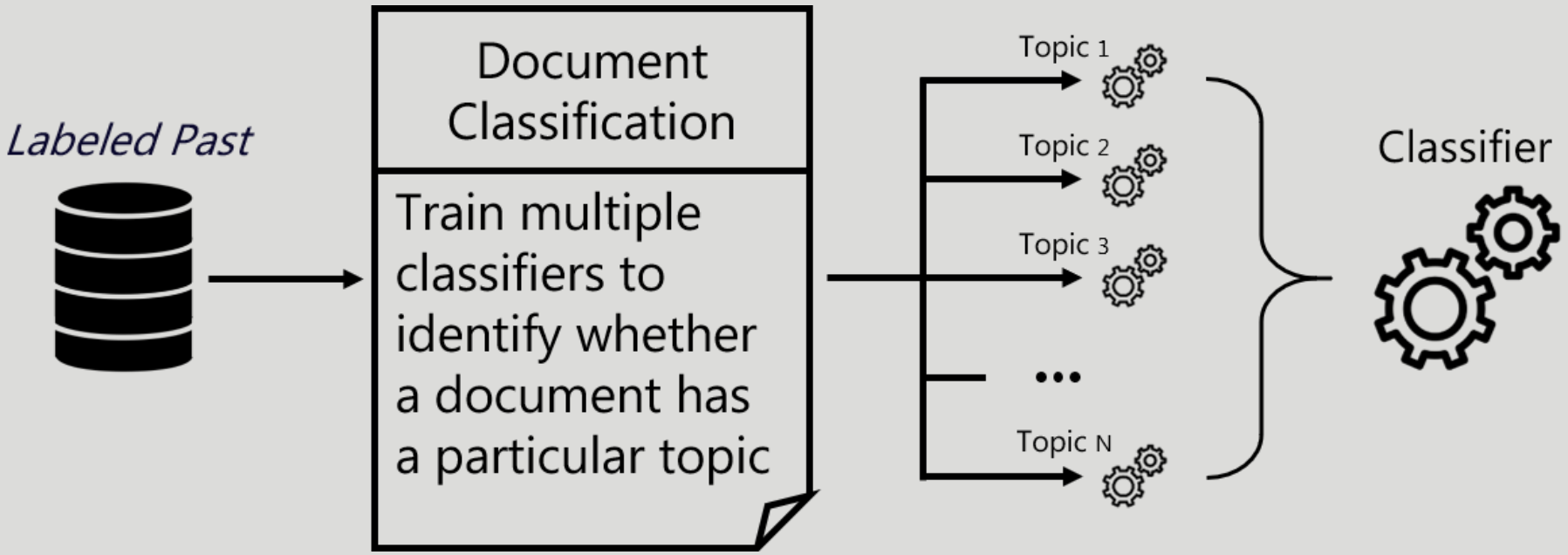


Present Topic 05

CONTENT

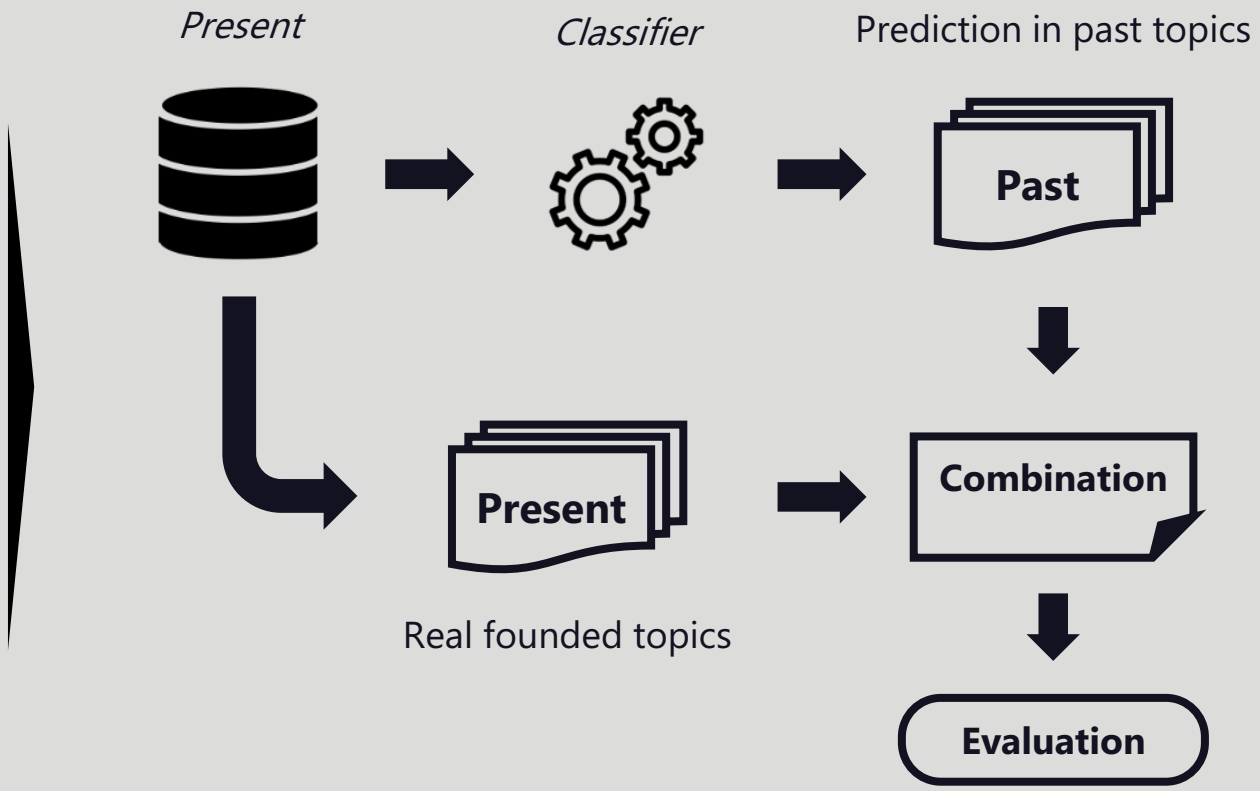
METHODOLOGY AND RESULTS |

For each topic in our set of discovered topics, we must be able to identify which topics are covered by a new document. To perform this, we will build an individual binary classifier for each topic.

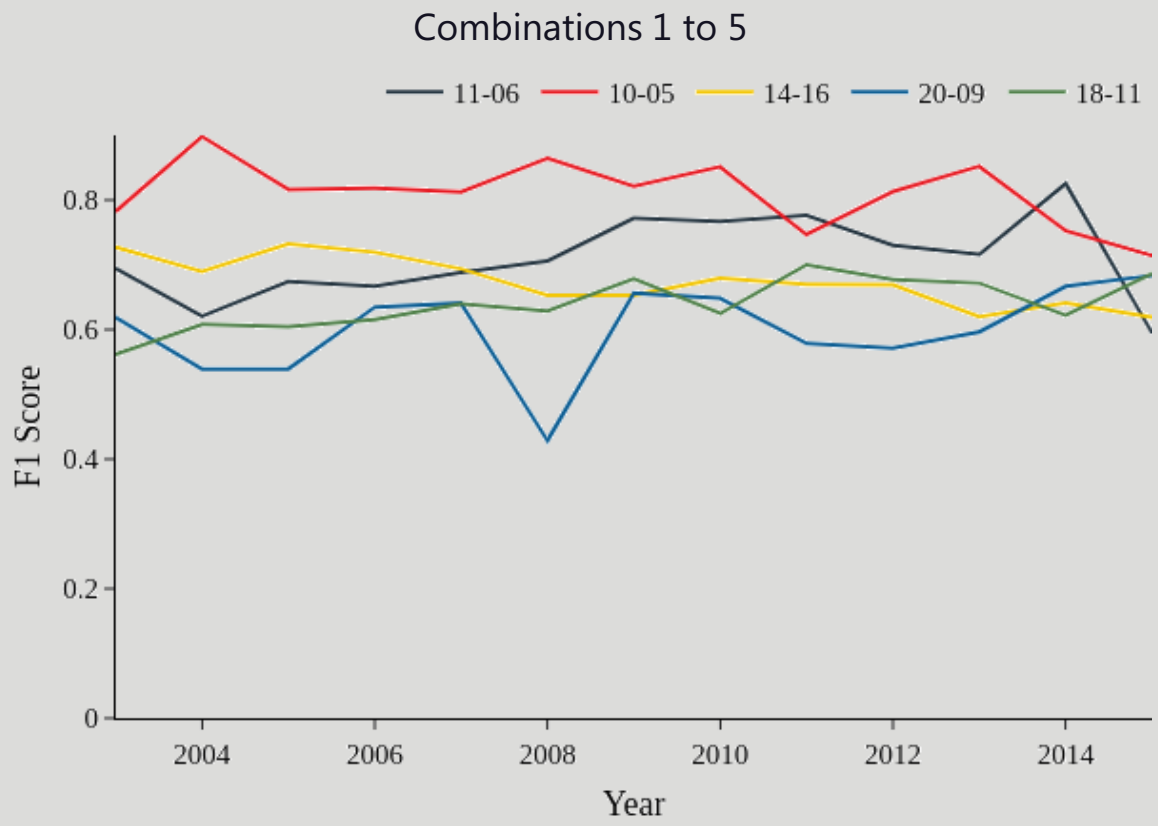


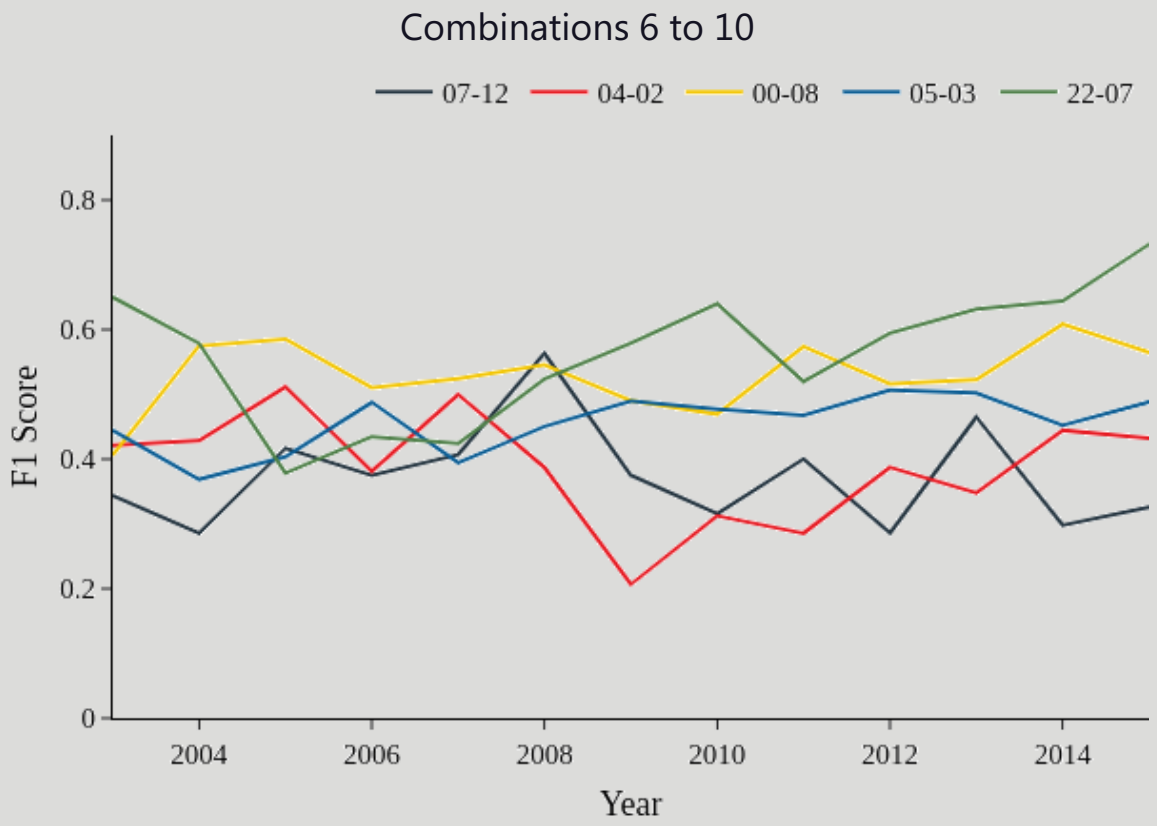
Model Selection

- 10-fold cross validation
- Combination
  - Vectorizer: TF-IDF
  - Model: SVM
- Evaluation metrics (mean)
  - Precision: 0.906
  - Recall: 0.683
  - F1 Score: 0.769



METHODOLOGY AND RESULTS |

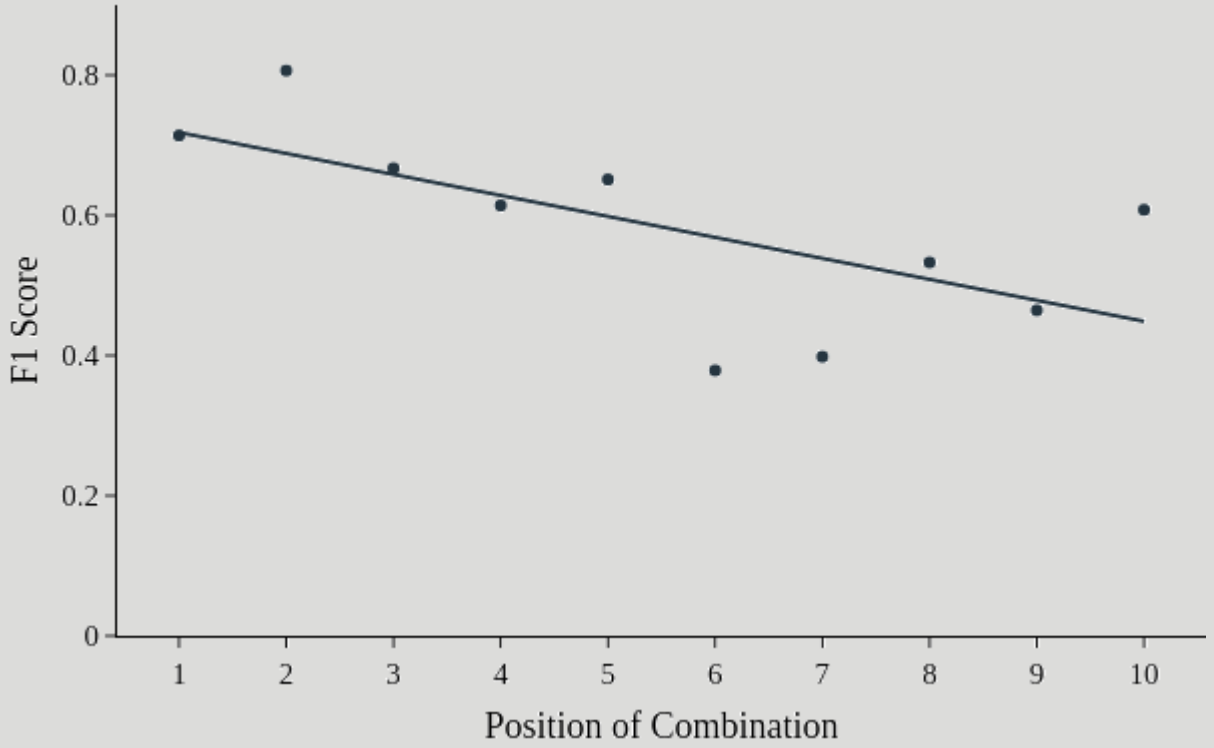




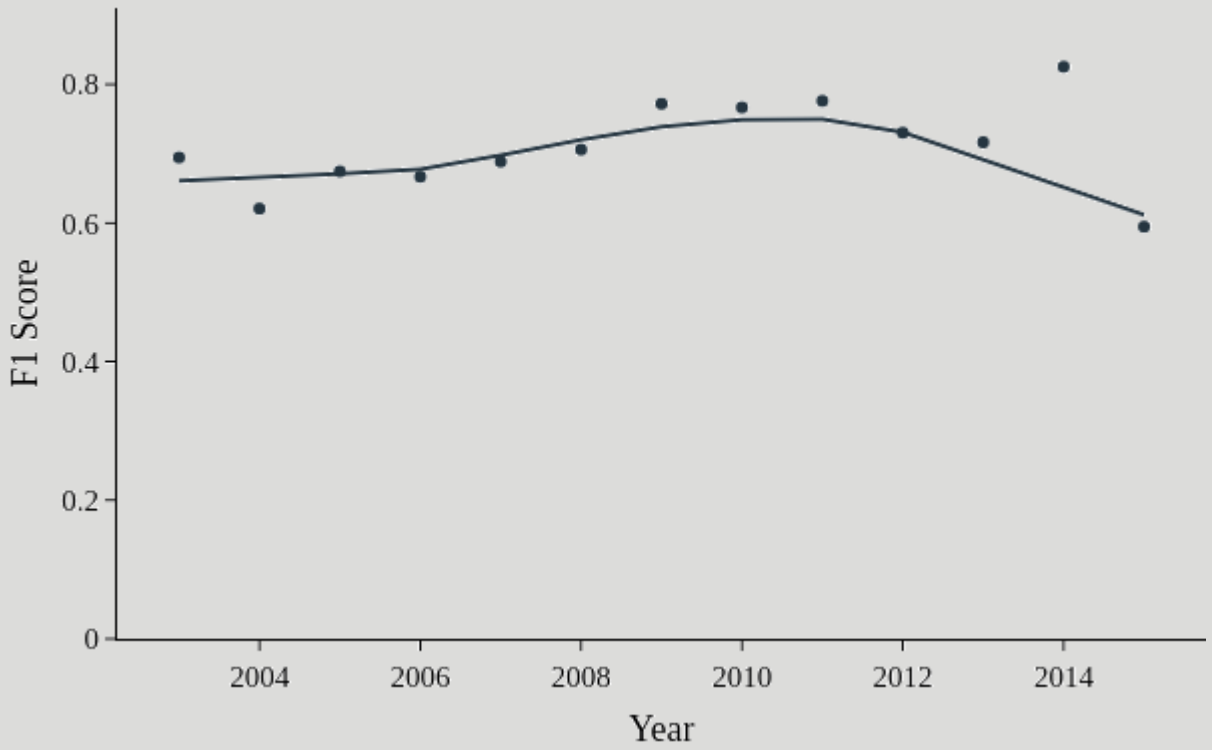
F1 Score *vs.* Position

- Negative correlation
  - -0.775
- Decreasing relationship
- Indicative for topics' evolution
  - New relevant terms
  - Emergence of new topics

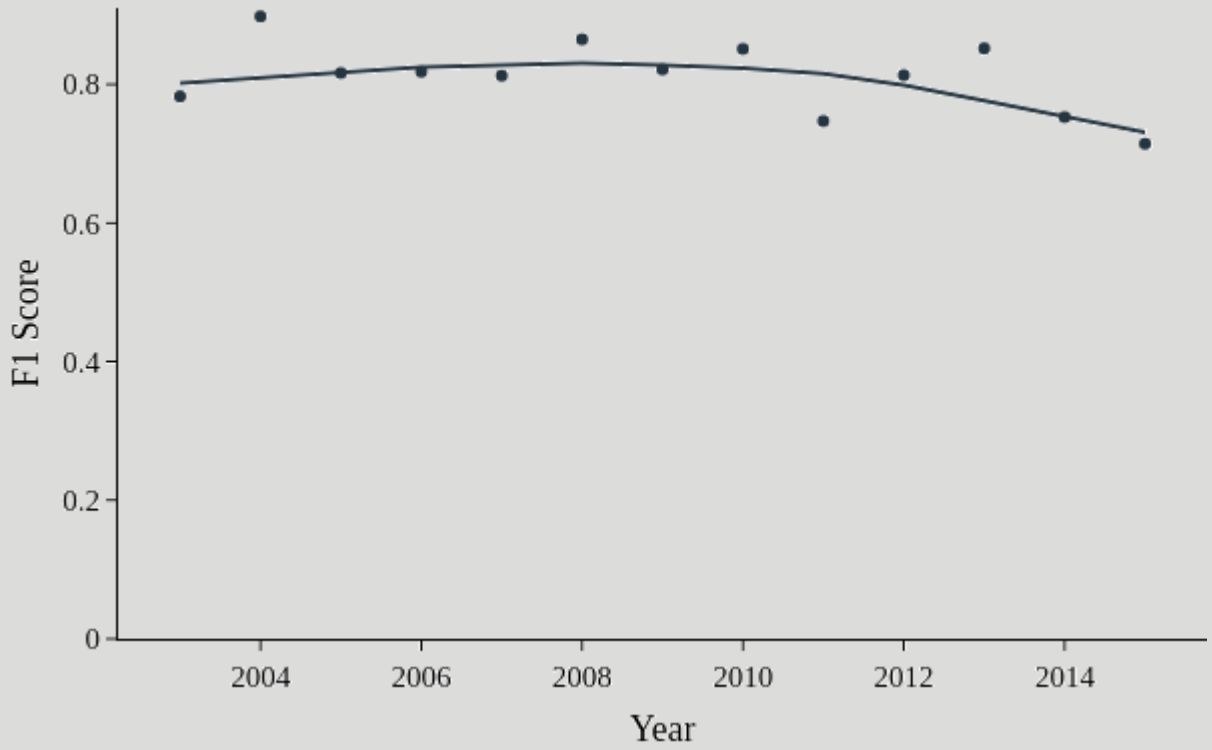
F1 Score correlation with the combination position



Yearly scoring for combination (*Past* 11 – *Pres* 06)



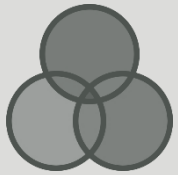
Yearly scoring for combination (*Past* 10 – *Pres* 05)



CONTENT

CONCLUSIONS

The main goal of this work was to evaluate the ability of classification models built from the topic modeling data, obtained by LDA. With *Past* subset, we obtained their natural topics, then we do the same for *Present*. After discovering the best combinations for them, we built classification models for *Past* topics and evaluate the predictions over the years.



Combination Efficiency

The lack of good combinations can indicate the evolution of the topics, either in the most relevant words to describe them or in the appearance of new topics

Yearly Evaluation

Constant tendency for topics that little develops, such as the 10-05 combination (reinforcement learning) followed by a slight drop in performance due to evolution

Future Works

- Forecast Step
  - Next framework step
  - Time series forecasting
- Retraining Point
  - Explore other granularity level
  - Identify efficacy loss for classification
- Feature extraction improvement
  - More improved techniques

Thank you!