

# Papers, pls!

A similarity-based recommendation  
system of research papers

**Aline Quadros**

**DATA SCIENCE**

**RETREAT**

**With more than 2.5  
million papers published  
per year...**

**How do we  
find relevant  
information  
for our  
research?**



PubMed

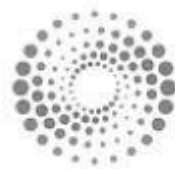
Scopus®

Google  
Scholar

DOAJ  
DIRECTORY OF  
OPEN ACCESS  
JOURNALS

SciELO

WorldCat®



WEB OF SCIENCE

1.

Keyword-based rankings  
can be influenced by  
changing the way you  
write the text

Statistical **natural language processing** «统计自然语言处理 »

X Zhang - 2014 - JSTOR

... As an encyclopaedic resource on statistical approaches to **natural language processing (NLP)**, the content of the second edition of Zong's (2008) **Statistical Natural Language Processing** (henceforth SNLP2) has some overlap with classic **NLP** textbooks such as Foundations of ...

☆ 99 Cited by 104 Related articles

1.

Keyword-based rankings  
can be influenced by  
changing the way you  
write the text

2.

Ranks now also  
rely on citations  
and other metrics

The screenshot shows a Google Scholar search interface. The search bar contains the text "natural language processing nlp". Below the search bar, the results are displayed. The first result is "Statistical natural language processing «统计自然语言处理»" by X Zhang, published in 2014 on JSTOR. The abstract mentions an encyclopaedic resource on statistical approaches to NLP, specifically referencing the second edition of Zong's (2008) "Statistical Natural Language Processing" (henceforth SNLP2). The citation count "Cited by 104" is circled in red. The second result is "[PDF] Natural language processing and its future in medicine" by C Friedman and G Hripcsak, published in 1999 at columbia.edu. The abstract discusses understanding NLP through syntax and XML. The citation count "Cited by 183" is also visible. The third result is "[PDF] Jumping NLP curves: A review of natural language processing research" by E Cambria and B White, published in 2014 in the IEEE Computational Intelligence Magazine. The abstract describes NLP as a theory-motivated range of computational techniques. The citation count "Cited by 561" is visible. The left sidebar contains filters for "Any time", "Since 2019", "Since 2018", "Since 2015", and "Custom range...". It also has sorting options "Sort by relevance" and "Sort by date". Checkboxes for "include patents" and "include citations" are present, with "include citations" checked. A "Create alert" button is at the bottom of the sidebar.

natural language processing nlp

scholar.google.com/scholar?hl=en&as\_sdt=0%2C5&q=natural+language+processing+nlp&oq=NLP

Google Scholar

natural language processing nlp

Articles About 130.000 results (0,07 sec)

Any time  
Since 2019  
Since 2018  
Since 2015  
Custom range...

Sort by relevance  
Sort by date

☒ include patents  
☒ include citations

☒ Create alert

Statistical **natural language processing** «统计自然语言处理»  
X Zhang - 2014 - JSTOR  
... As an encyclopaedic resource on statistical approaches to **natural language processing (NLP)**, the content of the second edition of Zong's (2008) **Statistical Natural Language Processing** (henceforth SNLP2) has some overlap with classic **NLP** textbooks such as Foundations of ...  
☆ 99 Cited by 104 Related articles

[PDF] **Natural language processing** and its future in medicine  
C Friedman, G Hripcsak - Acad Med, 1999 - columbia.edu  
... Understanding **natural language** involves understanding (1) syntax, or the structure of sentences (which ... 25 suggested using Extensible Markup Language (XML) to represent the **processed** output because ... String Project (LSP), is a pioneer both in **language processing** and in ...  
☆ 99 Cited by 183 Related articles All 6 versions

[PDF] Jumping **NLP** curves: A review of **natural language processing** research  
E Cambria, B White - IEEE Computational intelligence magazine, 2014 - krchowdhary.com  
48 IEEE Computational Intelligence Magazine May 2014 1556-603x/14/\$31.00©  
2014IEEE **natural language processing (NLP)** is a theory-motivated range of computational techniques for the automatic analysis and representation of human **language**. **NLP** research ...  
☆ 99 Cited by 561 Related articles All 11 versions

**Ranking of results is a challenging problem:  
Is there a more impartial way?**

**Ranking of results is a challenging problem:**  
**Is there a more impartial way?**

Can we use NLP to recommend papers based  
on text similarity?



**Ranking of results is a challenging problem:**  
**Is there a more impartial way?**

Can we use NLP to recommend papers based  
on text similarity?



**But how do we measure similarity?**  
**Similarity is also a challenging problem:**  
**Subjective task**  
**Unsupervised task**



# Dataset

# Process outline

**33510** papers  
(ID, title, and abstract)

from **arXiv.org**

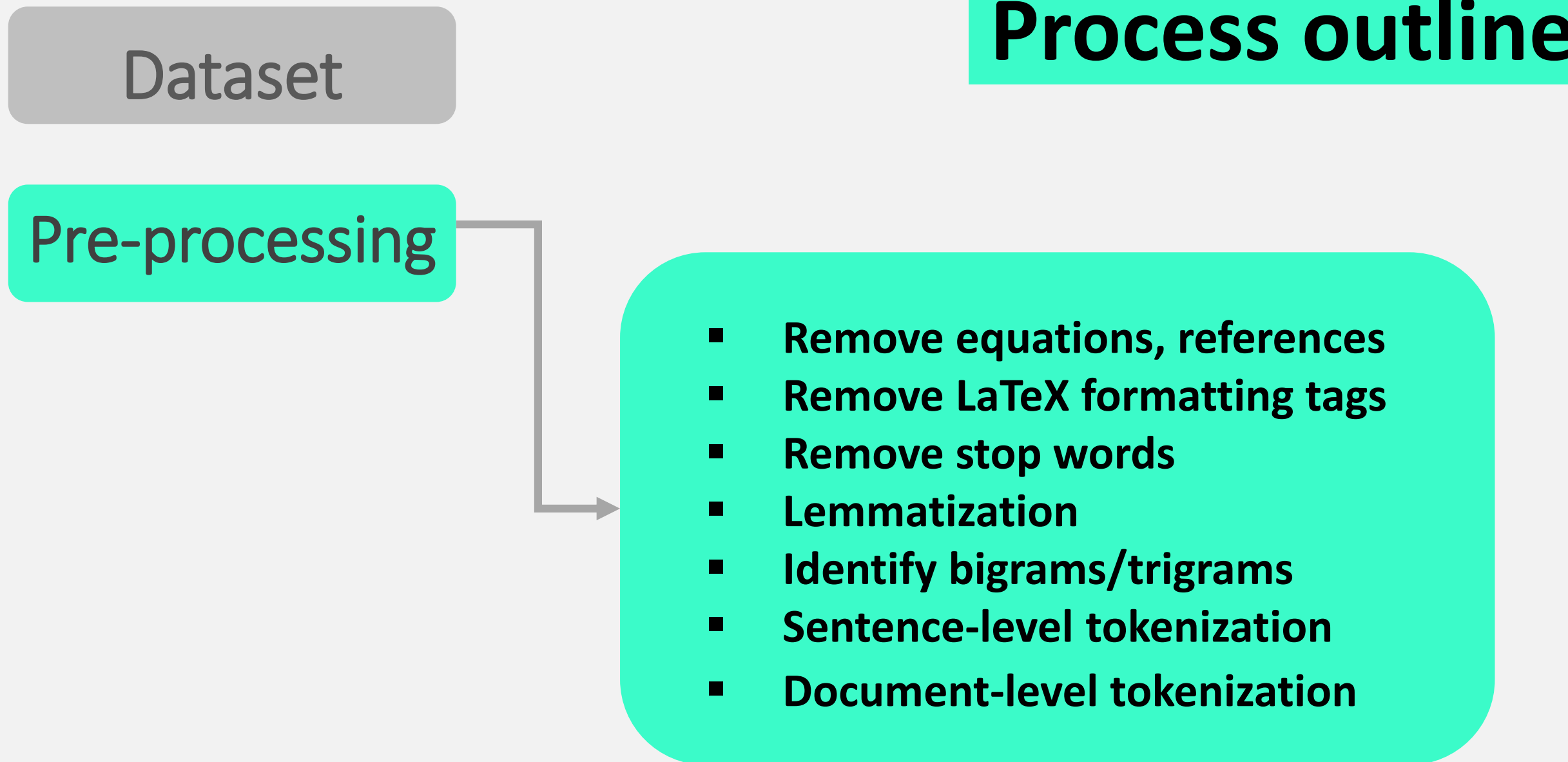
keyword  
**[machine learning]**



# Process outline

Dataset

Pre-processing

- 
- ```
graph LR; Dataset[Dataset] --> Pre-processing[Pre-processing]; Pre-processing --> Steps[Pre-processing steps];
```
- Remove equations, references
  - Remove LaTeX formatting tags
  - Remove stop words
  - Lemmatization
  - Identify bigrams/trigrams
  - Sentence-level tokenization
  - Document-level tokenization

# Process outline

Dataset

Pre-processing

Topic modelling



- Latent Dirichlet Allocation
- Metrics to find the “optimal” number of topics:
  - Coherence
  - Topics/document
  - Documents/topic
- Concatenated title + abstract

30

# TOPICS

## Broad topics

6k to 9k papers

[Learn,  
Propose,  
Method,  
Approach,  
Information,  
Data,  
Prediction]

## Well-defined topics

200 to 700 papers

[Bayesian,  
Likelihood,  
Inference,  
Markov]

## Very specific topics

10 to 60 papers

[Adversarial,  
Attack,  
Robust,  
Adversary,  
Threat]

# Process outline

Dataset

Pre-processing

Topic modelling



- Latent Dirichlet Allocation
- Metrics to find the “optimal” number of topics:
  - Coherence
  - Topics/document
  - Documents/topic
- Concatenated title + abstract

30

# Process outline

Dataset

Pre-processing

Topic modelling

**Embeddings**

## Paragraph Embeddings (Doc2Vec)

Pairwise cosine similarity between vector representations (embeddings) of each document

Distributed Memory version

The cat jumped on the sofa



The diagram illustrates the Distributed Memory version of Doc2Vec. It shows the sentence "The cat jumped on the sofa" with brackets underneath "The cat" and "on the sofa", indicating that these phrases are used to predict the missing word "jumped".

**Input:**

# Input:

<http://arxiv.org/abs/1903.00553v2>

## Attacking graph-based classification via manipulating the graph structure

Graph-based classification methods are widely used for security and privacy analytics. roughly speaking, graph-based classification methods include collective classification and graph neural network. evading a graph-based classification method enables an attacker to evade detection in security analytics and can be used as a privacy defense against inference attacks. existing adversarial machine learning studies mainly focused on machine learning for non-graph data. only a few recent studies touched adversarial graph-based classification methods. however, they focused on graph neural network methods, leaving adversarial collective classification largely unexplored...



## Input:

<http://arxiv.org/abs/1903.00553v2>

### Attacking graph-based classification via manipulating the graph structure

Graph-based classification methods are widely used for security and privacy analytics. roughly speaking, graph-based classification methods include collective classification and graph neural network. evading a graph-based classification method enables an attacker to evade detection in security analytics and can be used as a privacy defense against inference attacks. existing adversarial machine learning studies mainly focused on machine learning for non-graph data. only a few recent studies touched adversarial graph-based classification methods. however, they focused on graph neural network methods, leaving adversarial collective classification largely unexplored...

## Recommended:

<http://arxiv.org/abs/1805.07984v3>

### Adversarial attacks on neural networks for graph data

Deep learning models for graphs have achieved strong performance for the task of node classification. Despite their proliferation, currently there is no study of their robustness to adversarial attacks. yet, in domains where they are likely to be used, e.g. the web, adversaries are common. Can deep learning models for graphs be easily fooled? in this work, we introduce the first study of adversarial attacks on attributed graphs, specifically focusing on models exploiting ideas of graph convolutions. in addition to attacks at test time, we tackle the more challenging class of poisoning-causative attacks...

## Input:

### Bayesian optimization for dynamic problems

We propose practical extensions to bayesian optimization for solving dynamic problems. we model dynamic objective functions using spatiotemporal gaussian process priors which capture all the instances of the functions over time. our extensions to bayesian optimization use the information learnt from this model to guide the tracking of a temporally evolving minimum. by exploiting temporal correlations, the proposed method also determines when to make evaluations, how fast to make those evaluations, and it induces an appropriate budget of steps based on the available information. lastly, we evaluate our technique on synthetic and real-world problems.

## Recommended:

### Batched high-dimensional bayesian optimization via structural kernel learning

Optimization of high-dimensional black-box functions is an extremely challenging problem. while bayesian optimization has emerged as a popular approach for optimizing black-box functions, its applicability has been limited to low-dimensional problems due to its computational and statistical challenges arising from high-dimensional settings. [...] performing multiple evaluations in parallel to reduce the number of iterations required by the method. our novel approach learns the latent structure with gibbs sampling and constructs batched queries using determinantal point processes. experimental validations on both synthetic and real-world functions ...

## Input:

### An Empirical-Bayes Score for Discrete Bayesian Networks

Bayesian network structure learning is often performed in a Bayesian setting, by evaluating candidate structures using their posterior probabilities for a given data set. Score-based algorithms then use those posterior probabilities as an objective function and return the maximum a posteriori network as the learned model. For discrete Bayesian networks, the canonical choice for a posterior score is the Bayesian Dirichlet equivalent uniform (BDeu) marginal likelihood with a uniform (U) graph prior (Heckerman et al., 1995). Its favourable theoretical properties descend from assuming a uniform prior both on the space of the network structures and on the space of the parameters of the network. In this paper, we revisit the limitations of these assumptions; and we introduce an alternative set of assumptions and the resulting score: the Bayesian Dirichlet sparse (BDs) empirical Bayes marginal likelihood with a marginal uniform (MU) graph prior. We evaluate its performance in an extensive simulation study, showing that MU+BDs is more accurate than U+BDeu both in learning the structure of the network and in predicting new observations, while not being computationally more complex to estimate.

## Recommended:

### Beyond Uniform Priors in Bayesian Network Structure Learning

Bayesian network structure learning is often performed in a Bayesian setting, evaluating candidate structures using their posterior probabilities for a given data set. Score-based algorithms then use those posterior probabilities as an objective function and return the maximum a posteriori network as the learned model. For discrete Bayesian networks, the canonical choice for a posterior score is the Bayesian Dirichlet equivalent uniform (BDeu) marginal likelihood with a uniform (U) graph prior, which assumes a uniform prior both on the network structures and on the parameters of the networks. In this paper, we investigate the problems arising from these assumptions, focusing on those caused by small sample sizes and sparse data. We then propose an alternative posterior score: the Bayesian Dirichlet sparse (BDs) marginal likelihood with a marginal uniform (MU) graph prior. Like U+BDeu, MU+BDs does not require any prior information on the probabilistic structure of the data and can be used as a replacement noninformative score. We study its theoretical properties and we evaluate its performance in an extensive simulation study, showing that MU+BDs is both more accurate than U+BDeu in learning the structure of the network ...

# Conclusions

**LDA improved with combination of title + abstract**

**Doc2Vec and LDA can be used to detect similar papers**

**They are more powerful together**

**Doc2Vec identifies plagiarism/copies**

**Dataset needs to be larger**

The data says we need more data.



someecards  
user card

## TODO list:

- **Explore pre-trained sub-word models**
- **Enhance the use of Universal Sentence Encoder**
- **Expand the dataset coverage**

## TODO list:

- Explore pre-trained sub-word models
- Enhance the use of Universal Sentence Encoder
- Expand the dataset coverage

**Thank you!**

**Let's connect:**

**Aline.fquadros@outlook.com**



alinequadros

## TODO list:

- Explore pre-trained sub-word models
- Enhance the use of Universal Sentence Encoder
- Expand the dataset coverage

**Thank you!**

**Let's connect:**

**Aline.fquadros@outlook.com**



alinequadros



@alinequadros