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®

JSTOR

BMC

WEB OF SCIENCE

1.

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

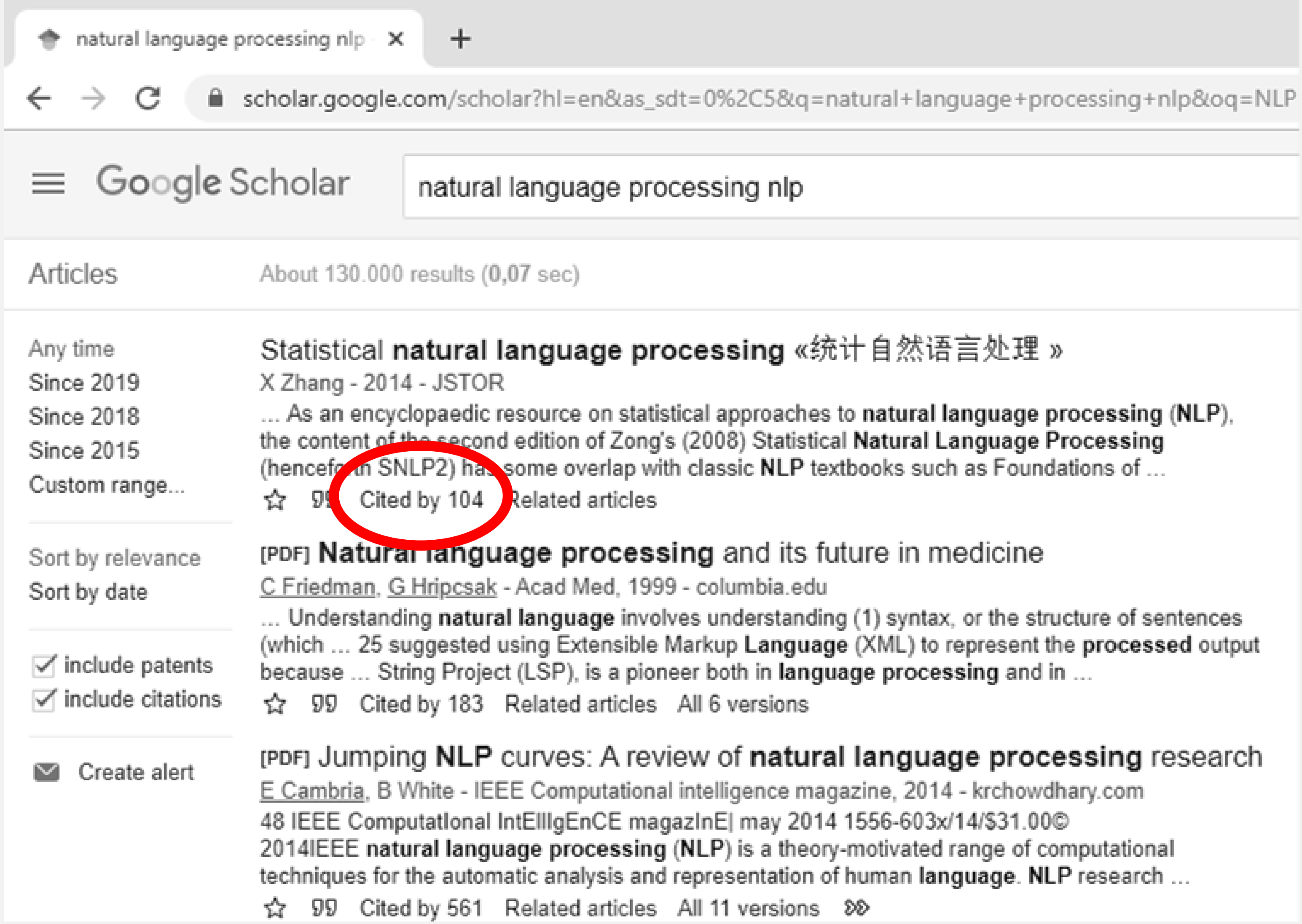
Statistical **natural language processing** «统计自然语言处理 »
X Zhang - 2014 - JSTOR
... As an encyclopaedic resource on statistical approaches to **natural language processing** (**NLP**), the content of the second edition of Zong's (2008) Statistical **Natural Language Processing** (henceforth SNLP2) has 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



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

Process outline

Dataset

33510 papers

(ID, title, and abstract)

from **arXiv.org**

keyword

[machine learning]

Process outline

Dataset

Pre-processing

- 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**

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

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