CS 5/7391 Spring 2021

Project

The goal of the project is to allow student to explore and develop various problems in Natural Language Processing.

I will provide a list of project below. You should rank the projects in order of preference and e-mail me back your options by 5pm, 3/29 (Mon). You are welcomed to form a team of 2 and submit your order as a group. Eventually I will put 2-3 people in a project, and people for each project may collaborate or work on different aspect of it. The project assignment will be announced on Monday's class.

Each project (and at each level) will have its own required set of deliverables, which include code and write-ups. (Programming will be done mostly in Python, but there are projects where Java or other languages are preferable).

For each project there is the following steps:

- I will meet each project group on a weekly basis in April. The meeting will roughly be 10-15 minutes. I will work with each group for a time slot, but it will be either Thursday (daytime), or Friday (daytime or evening). For each meeting (starting the second one) there will be milestones I expect each group to finish by then. Overall progress through the milestones will count towards 25% of the project grade.
- For each project there will be 1-2 paper(s) associated with it that provide some background information. Each group will give a 25-minute presentation of the paper on either 4/26 or 4/28. You can have either one person or multiple members of the group to present the papers. The presentation will count towards 10% of the project grade.
- Each group will need to present its work on 5/11 (Tue) between 3-6pm. Each group will have around 15-20 minutes for its presentation. More details will be provided latger. This will count towards 15% of the grade
- The final deliverables for each project need to be uploaded to Canvas (as a zip file) by noon 5/12 (Wed).

List of projects

1. Overcoming instability for LDA

   One of the limitation of LDA is that it is highly unstable – if you run the same LDA on the same data multiple times, they are going to return results that are different. In this project, our goal is to develop methods that overcome this problem by running LDA for multiple times and apply transformation to the results.

   Papers:

   - Maëlick Claes, Maelick Claes, Umar Farooq. *Measuring LDA topic stability from clusters of replicated runs,* ESEM '18: Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement

- Derek Greene, Derek O'Callaghan, Pádraig Cunningham, *How Many Topics? Stability Analysis for Topic Models.* CoRR abs/1404.4606 (2014)

Software modules:

- Gensim (LDA code): Gensim: Topic modelling for humans (radimrehurek.com)

Project 2 and 3 tries to solve a problem:

“We are given a set of text and a certain entity (people, company etc.) We want to discover as much information about that the object and represent it in the form of wikidata.”

Information about Wikidata

- https://www.wikidata.org
- Denny Vrandečić, Markus Krötzsch, *Wikidata: a free collaborative knowledgebase*, Communications of the ACM, 57 (10), Oct 2014.

2. Extracting information from text via openIE

We will use openIE as the basis of extracting information from text. The project will first run openIE “as is” to extract the info, and then try to put together the information together in the format of WikiData

Papers:

- Gabor Angeli, Melvin Johnson Premkumar and Christopher D. Manning. *Leveraging Linguistic Structure For Open Domain Information Extraction.* Association for Computational Linguistics (ACL).
- Mausam. *"Open Information Extraction Systems and Downstream Applications".* Invited Paper for Early Career Spotlight Track. International Joint Conference on Artificial Intelligence (IJCAI). New York, NY. July 2016.

Software modules:

- Stanford openIE

3. Do co-reference resolution improve information extraction?

Co-reference resolution problem is the problem of relating references within text documents. For instance, “John arrives here at 10am. He is carrying a box of books.”. Co-reference resolution’s goal is to find out that “He” is referring to John.

The goal of this part of the project is to apply co-reference resolution to text documents before applying the information extraction algorithm to see if there is any improvement.

Papers:

- Hongming Zhang, Xinran Zhao, Yangqiu Song, *A Brief Survey and Comparative Study of Recent Development of Pronoun Coreference Resolution,*
- Kevin Clark and Christopher D. Manning, Deep Reinforcement Learning for Mention-Ranking Coreference Models, EMNLP 2016.

Software Module

- Neural Coref on Spacy: https://spacy.io/universe/project/neuralcoref

4. Implementing and updating PLSI

You will be given an implementation of PLSI in python. The goal is to figure out the algorithm that the program uses and then try to modify to see if there is any improvement that can be made.

Papers:

- Ayman Farahat, Francine Chen*, Improving Probabilistic Latent Semantic Analysiswith Principal Component Analysis*, 11th Conference of the European Chapter of the Association for Computational Linguistic, 2006
- Michael Nokel, Natalia Loukachevitch, *A Method of Accounting Bigrams in Topic Models*, Proceedings of the 11th Workshop on Multiword Expressions, 2015

Software Module:

- PLSI (PLSA) in Python: https://pypi.org/project/plsa/

5. Pre-processing methods for LDA

This project explores any pre-processing steps that can be applied to documents to improve the performance of topic modelling such as LDA. This include term-weighting and using coreference resolution.

Papers:

- C. Truica, F. Radulescu and A. Boicea, *"Comparing Different Term Weighting Schemas for Topic Modeling,"* 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*,* Timisoara, Romania, 2016, pp. 307-310, doi: 10.1109/SYNASC.2016.055.
- Fiona Martin, Mark Johnson, *More efficient topic modelling through a noun only approach*, Proceedings of the Australasian Language Technology Association Workshop 2015

Software modules:

- Gensim (LDA code): Gensim: Topic modelling for humans (radimrehurek.com)
- Neural Coref on Spacy: https://spacy.io/universe/project/neuralcoref

6. Information extraction from Biomedical text

You are given a set of research papers from the biomedical field. Our goal is to extract specify information from those text, most importantly drug-drug and protein-drug interaction. These information are particular useful in help develop new drugs for certain disease (e.g. cure for COVID-19)

(Notice that you do NOT need any biomedical background for this project)

Papers:

- Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu, Chitta Baral, *IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text*, Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, June 2005
- Ashok Thillaisundaram, Theodosia Togia *Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture*, Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Nov 2019

Software Module

- Stanford openIE