NLP 053 - CSE UOI 2025 Kaggle Challenge

Citation prediction: build a model to predict whether a paper cites another paper

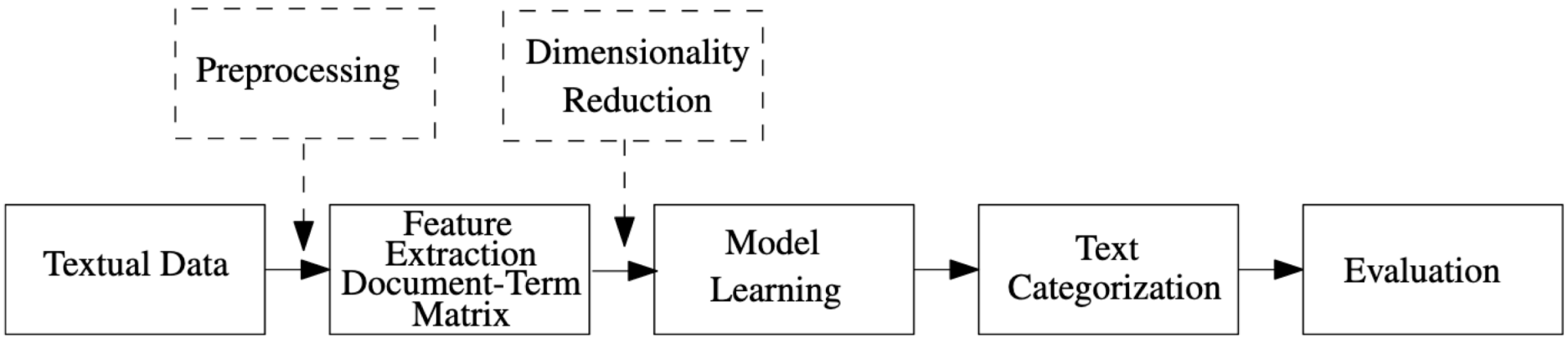
Citation Prediction (Link Prediction)

- Your task is to predict whether one research paper cites another based on a citation network of several thousand papers
- Each paper comes with an abstract, a list of authors, and information about which papers it currently cites
- By analyzing patterns within this citation data, you aim to build a model that forecasts the existence of a link (i.e., a citation) between two papers that do not already have one.
- Supervised learning pipeline: first, you extract features that capture the nature of connections between two papers—this could include textual similarity of abstracts, overlap or relatedness of authors, as well as network-based metrics (e.g., common neighbors or shortest path distance in the citation graph)
- After creating these feature vectors for paper pairs, you train a classifier on known citation links (positive examples) and known non-links (negative examples)
- Finally, you use this trained model to infer whether a new pair of papers is likely to be linked by a citation or not, effectively predicting future or currently undiscovered citations

Key points

- Deadline: 30-05-2024 (3 μήνες)
- Ομάδες: 1-2 άτομα
- 5 submissions per day
- Python, Ipython Notebook (+ ότι βιβλιοθήκη θέλετε)
- Report & presentation
  - 1 άτομο->min 4σελιδο, 2 άτομα-> min 5σελιδο (best->Latex)
  - Presentation: 15 slides max
- **Πρέπει να περάσετε τις προφορικές εξετάσεις για να μετρήσει η εργασία σας!**

What to show



- Να περιέχει ανάλυση δεδομένων
- Πώς αντιμετωπίσατε κάθε στάδιο του προβλήματος:
  - τι είδους αναπαράσταση δεδομένων χρησιμοποιήσατε, ποια χαρακτηριστικά χρησιμοποιήσατε,
  - αν εφαρμόσατε τεχνικές μείωσης των διαστάσεων
  - ποιους αλγορίθμους δοκιμάσατε και γιατί, την απόδοση των μεθόδων σας, χρόνος εκπαίδευσης
  - οποιεσδήποτε προσεγγίσεις που τελικά δεν λειτούργησαν αλλά είναι ενδιαφέρον να παρουσιάσετε, και γενικά, ό,τι νομίζετε ότι είναι ενδιαφέρον να αναφέρετε

Notes

- Link prediction (classification problem)
- Python packages: pandas, networkx, scikit-learn, gensim (for embeddings), tensorflow
- Analysis, visualization, feature selection
- Cross-validation to get score locally (not on Kaggle)
- Word2vec, Glove, Fasttext, BERT embeddings (for text)
- Node, edge features (centralities, link prediction algorithms, embeddings)
- Multiple classifiers and SOTA methods (after you get a good score)

Extra points

- Most creative team name
  - Computer science + news
- Fastest solutions
  - Before I upload the first baseline
- Running SOTA method
  - Be the only one who used an approach

Have fun!