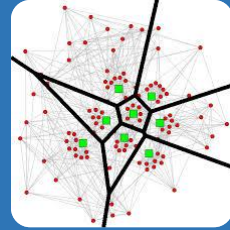


A Weakly Supervised Graph-based System for Customer Review Categorization

Eray Yıldız

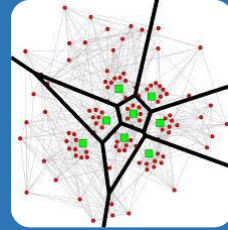
TERM PROJECT

for ITU BLG 614E - WEB MINING COURSE

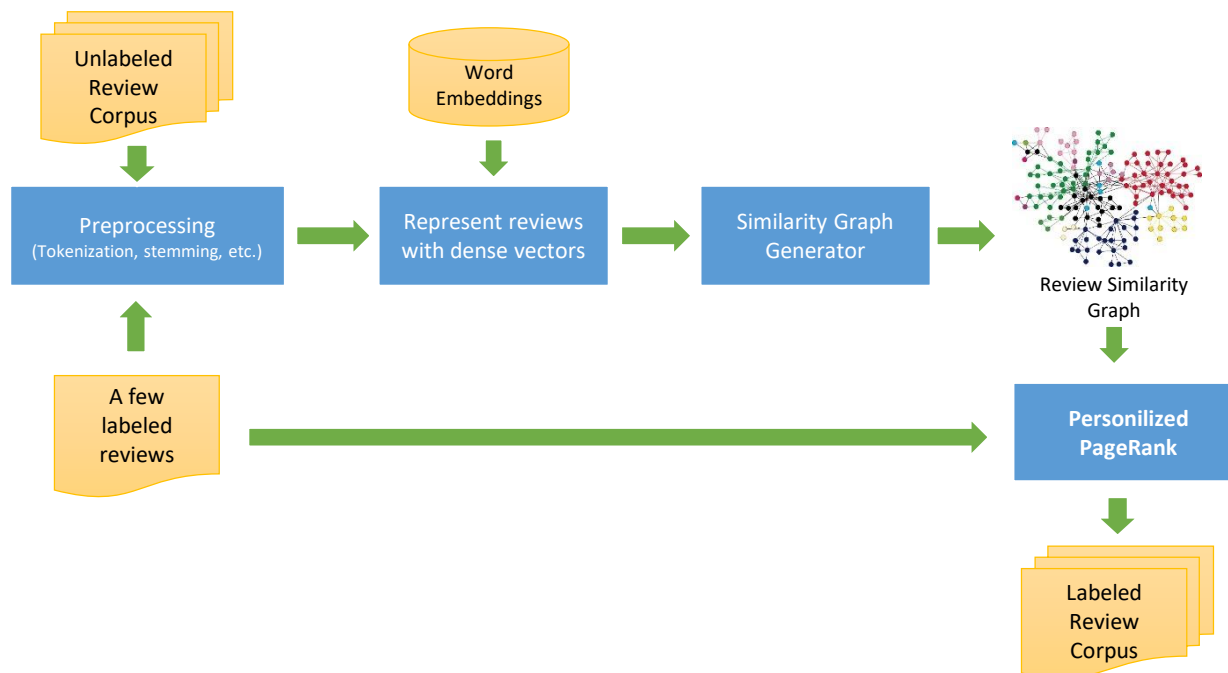


- Assigning aspect categories to customer reviews is an important problem
 - The providers want to monitor the opinion of their costumers with respect to a set of aspects such as **price**, **service** and **quality**
- Supervised learning methods are mostly used for aspect categorization
 - Although supervised systems obtain good results for the domain they are trained on, they are almost useless when the domain or language has changed.
 - Annotating data for all domains and languages is not applicable.
- **The goal** is to develop a semi-supervised system
 - which is able to automatically categorize customer reviews into aspect categories **defined by the users**.
 - which requires only a **small amount** of labeled reviews
- **Can Personalized PageRank Algorithm help?**

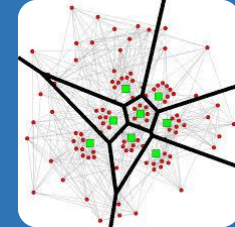
Proposed Method



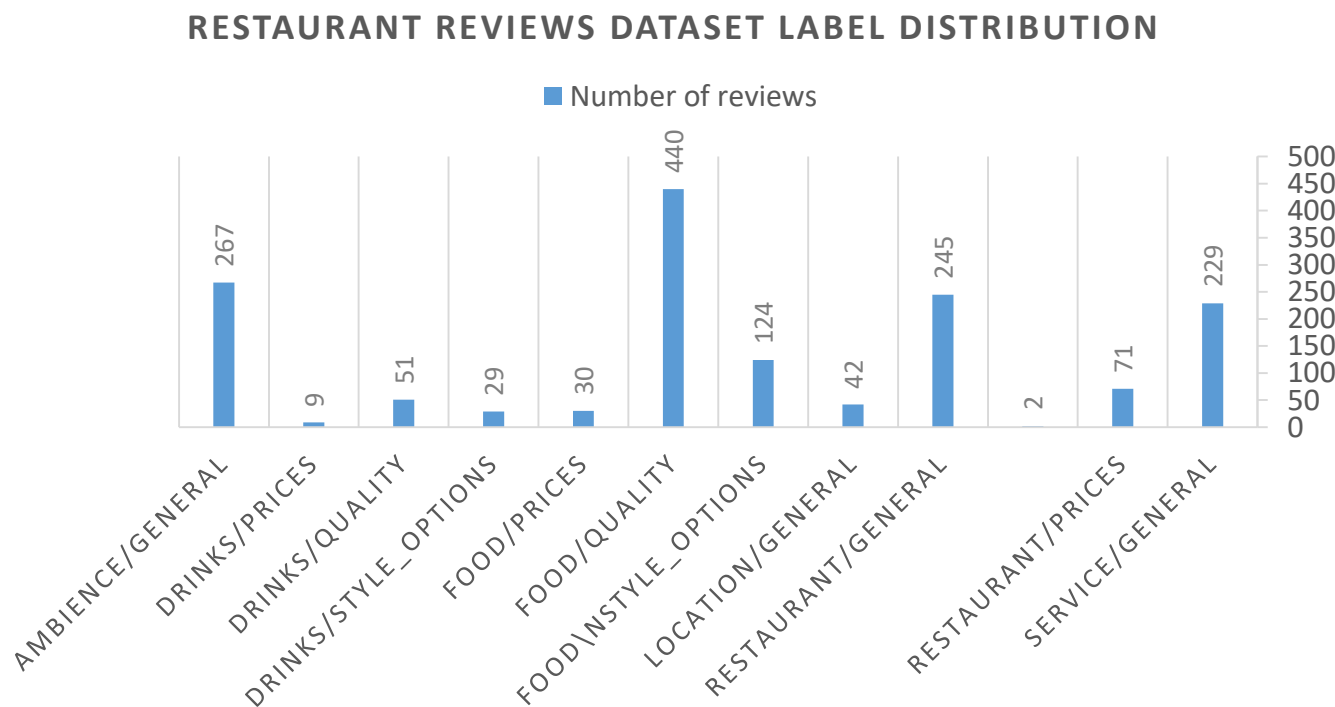
- Preprocess all reviews (e.g., tokenization and stemming)
- Represent each review with a dense vector using doc2vec
- Construct similarity graph using cosine similarity of review vectors
- Apply Personalized PageRank (labeled reviews as starting nodes of random walks)



Dataset used for evaluation

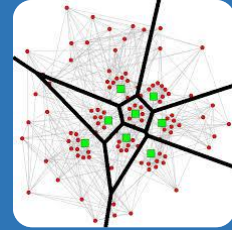


- Turkish restaurant reviews dataset
 - Published in SemEval 2016 Task 5 [1]
 - Total number of reviews: 1228
 - Each review has multiple labels (multi-label classification task)



[1] Pontiki, Maria, et al. "SemEval-2016 task 5: Aspect based sentiment analysis." *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 2016.

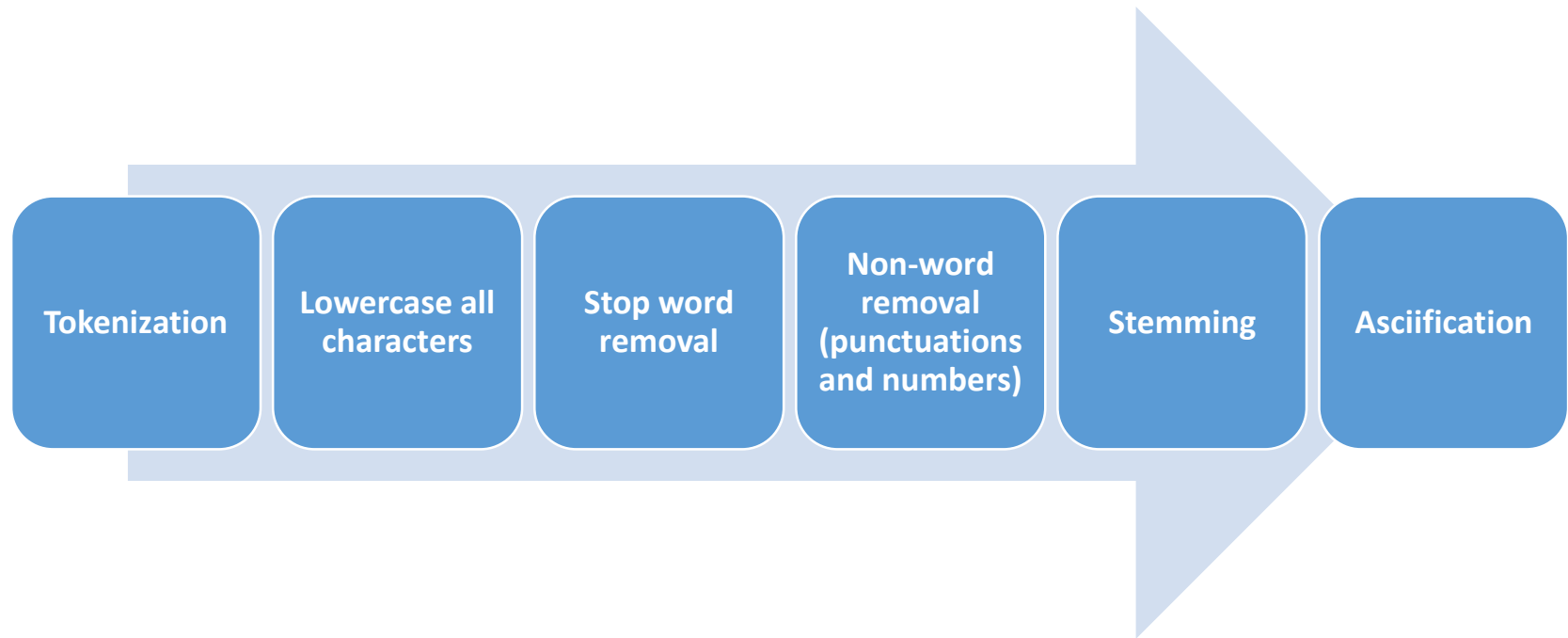
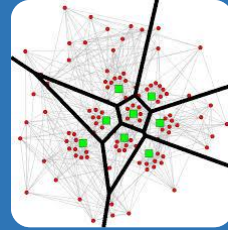
Dataset used for evaluation



Example reviews and labels from dataset

Review	Aspect categories
fiyatlar normal, servis cok yavaş, tuvaletler ise pis.	AMBIENCE/GENERAL SERVICE/GENERAL RESTAURANT/PRICES
bu servise bu fiyatlar ise cok fazla.	RESTAURANT/PRICES
1 kere denedigim fiyatlarinin yüksek yemeklerinin lezzetsiz oldugu bir mekan önermiyorum	RESTAURANT/PRICES FOOD/QUALITY
Duyduğum kadariyla suda değil sütte bekletiyorlarmış	FOOD#STYLE OPTIONS
Sevmediğimden yiyemedim kenarlarını, ama ortası pek lezzetli	FOOD/QUALITY FOOD/STYLE OPTIONS

Preprocessing steps:

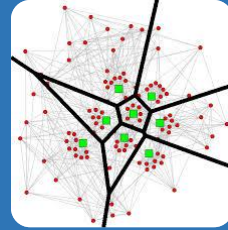


fiyatlar normal, servis cok
yavas, tuvaletler ise pis.



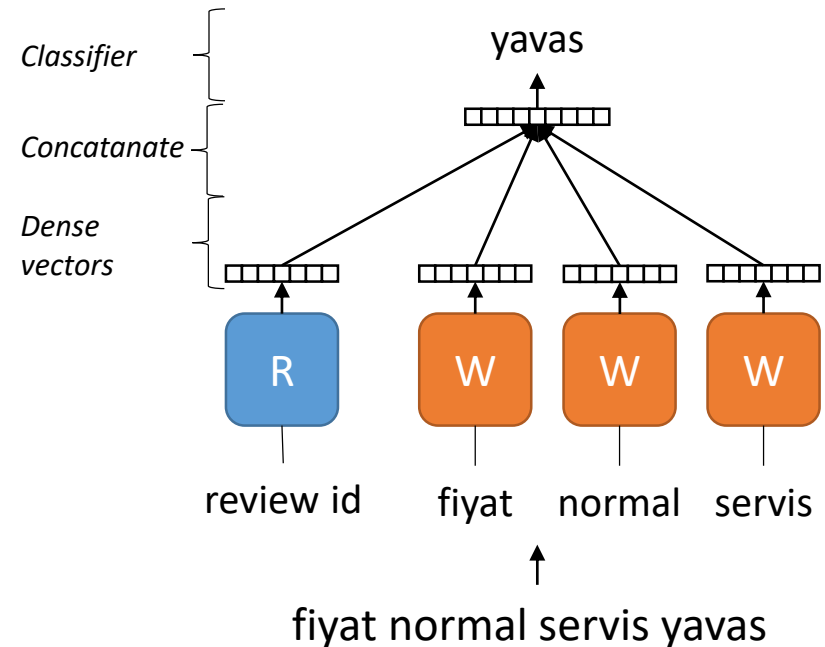
fiyat normal servis yavas
tuvalet pis

Learning vector representation of reviews



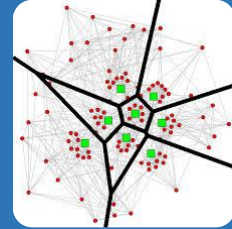
- Distributed Representations of Documents (Doc2Vec) is used to obtain vector representation of reviews
 - Similar to word2vec except it also generates vector representations of documents in addition to word vectors
- Doc2Vec model is trained on a large customer review dataset
 - 1.2 million reviews are obtained by crawling sources such as eksisozluk, hepsiburada and sikayetvar
 - Restaurant reviews are also added into the dataset

Distributed Representations of Documents[1] (a.k.a Doc2Vec)



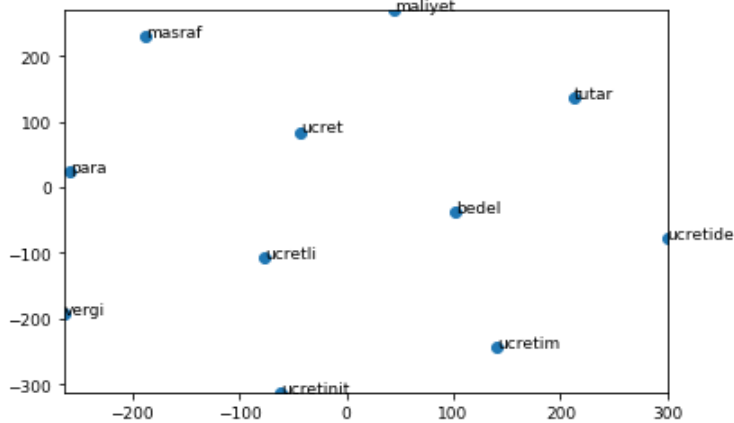
[2] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International Conference on Machine Learning*. 2014.

Learning vector representation of reviews

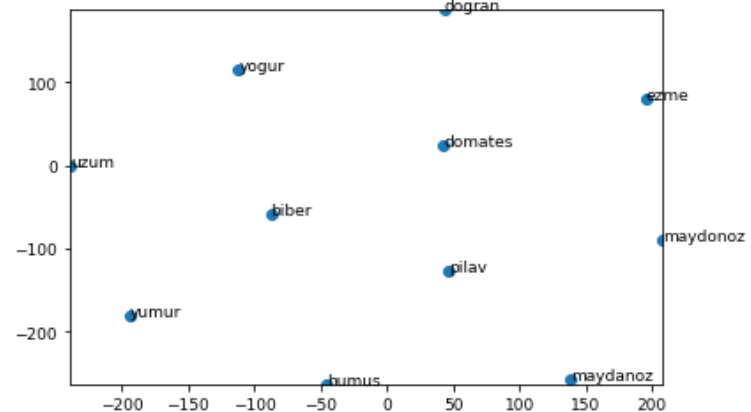


Some Word Vectors in two dimensional space
(dimension reduction is performed by TSNE algorithm)

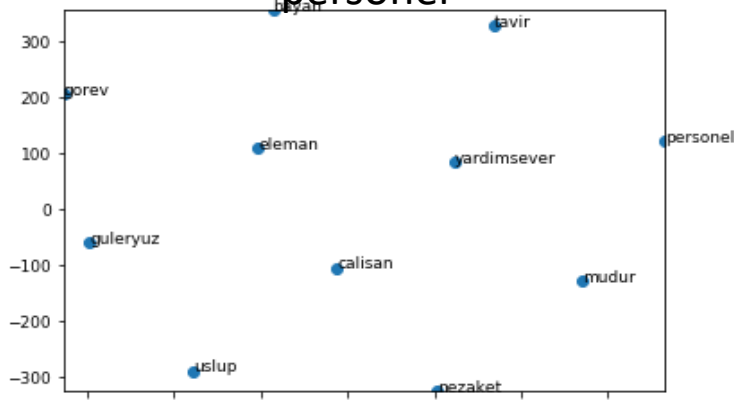
ücret



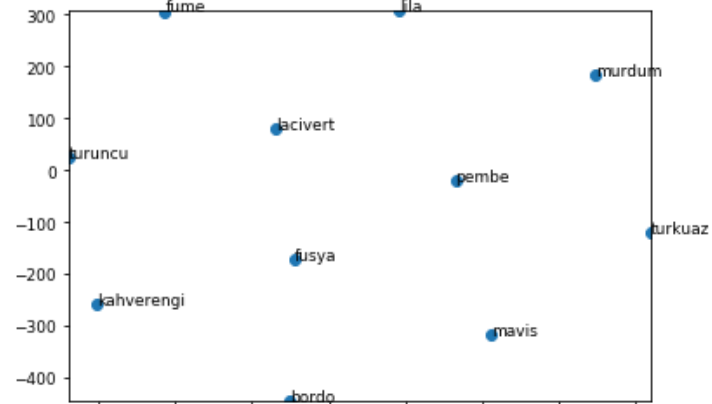
maydanoz



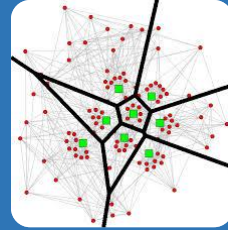
personel



turkuaz



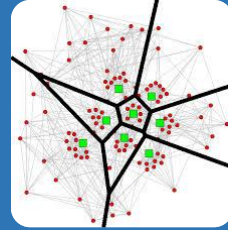
Learning vector representation of reviews



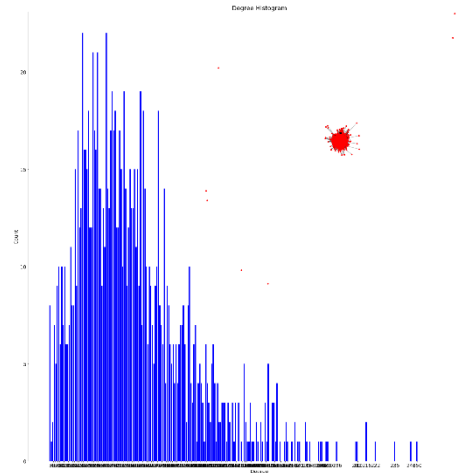
Some review vectors in two dimensional space
(dimension reduction is performed by TSNE algorithm)

- bir tabaga 5 farkli cesit isteyebilirsiniz, ozellikle bugday salatasi ve havuc sal
- menüdeki çeşitleri sorunca, o yok, bu yok, onun yerine şunu verelim, bunu verelim.
- özellikle bahçesini tercih edin giderseniz.
- fena değil ama menemen için i
- ambiansi hos.

Construction of Similarity Graphs

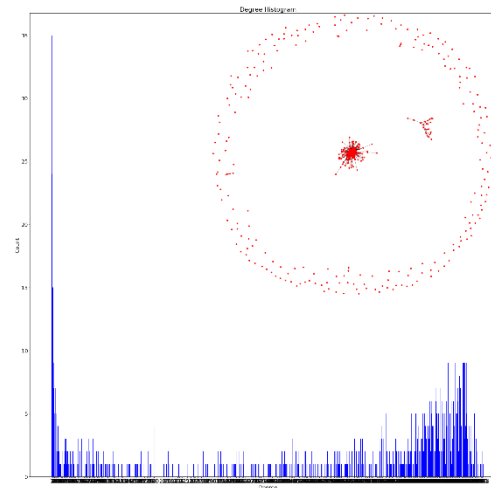


- Calculate cosine similarity between each review pair using vector representations
 - Both for Doc2Vec and TF-IDF vectors



Degree distribution of TF-IDF similarity graph

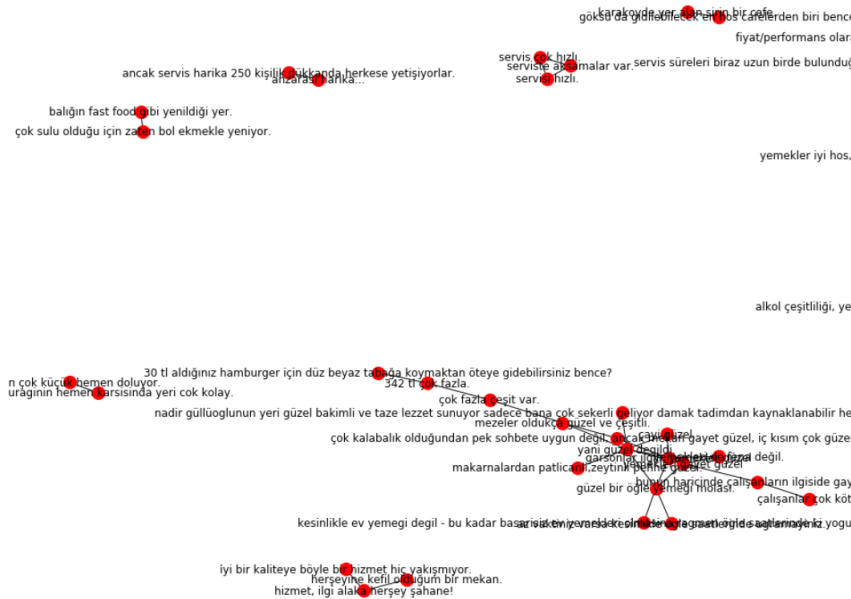
Threshold: 0.1
Average degree count = 59
Nodes without connection = 8



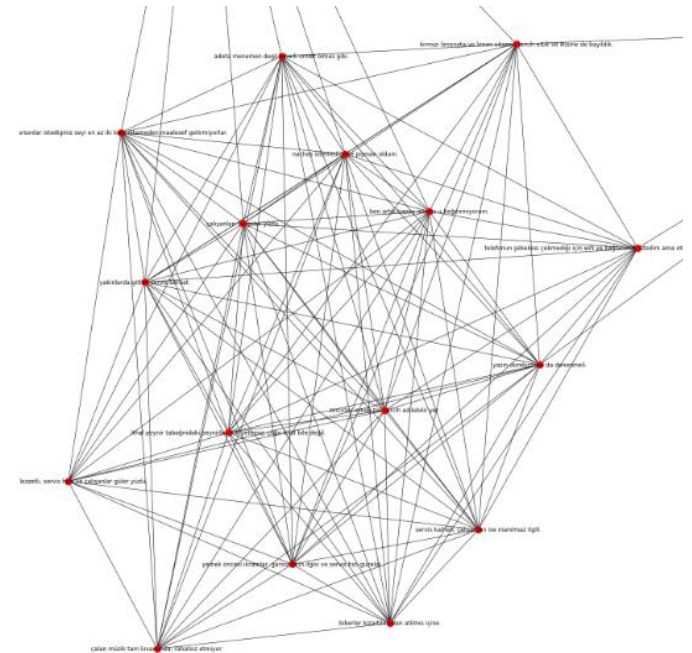
Degree distribution of DOC2VEC similarity graph

Threshold: 0.7
Average degree count = 498
Nodes without connection = 226

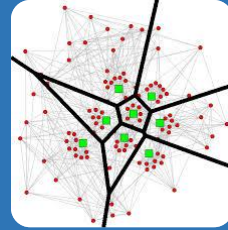
A subgraph of TF-IDF similarity graph



A subgraph of Doc2Vec similarity graph



Personalized PageRank for Multi-label Categorization



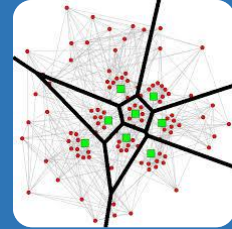
- Personalized PageRank (PPR) is modified version of PageRank
- Except jumps are back to one of a given set of starting vertices.
 - with probability α , follow a random edge
 - with probability $1 - \alpha$ restart at a seed

Idea: (PPR for multi-label categorization)

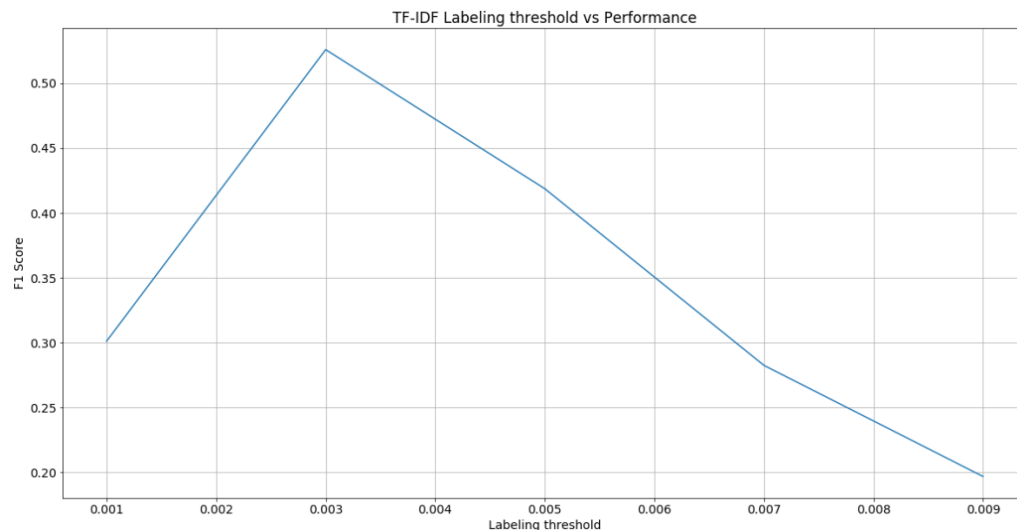
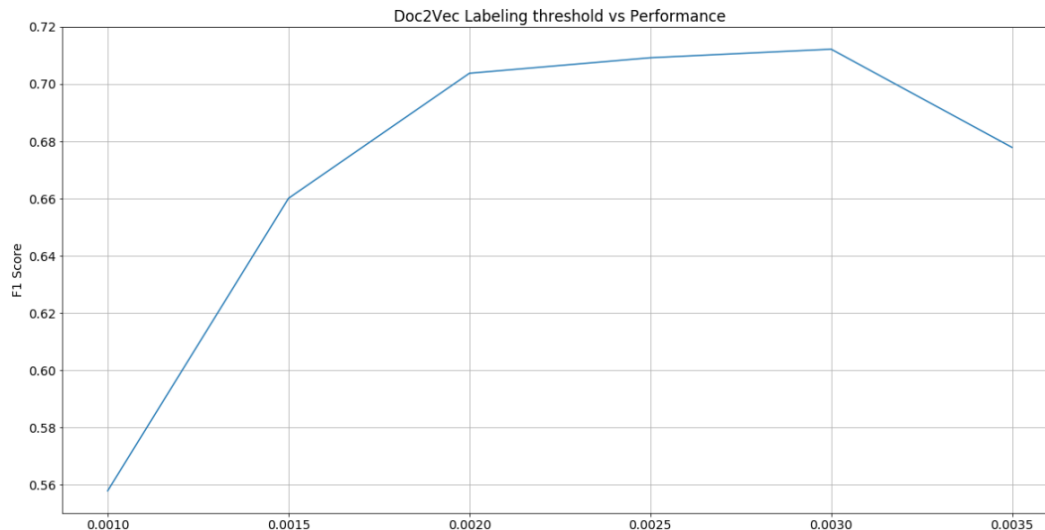
- For each label in the dataset:
 - Use labeled reviews as seeds for PPR
 - Apply a threshold to PPR scores of the nodes

PPR for Multi-label Categorization

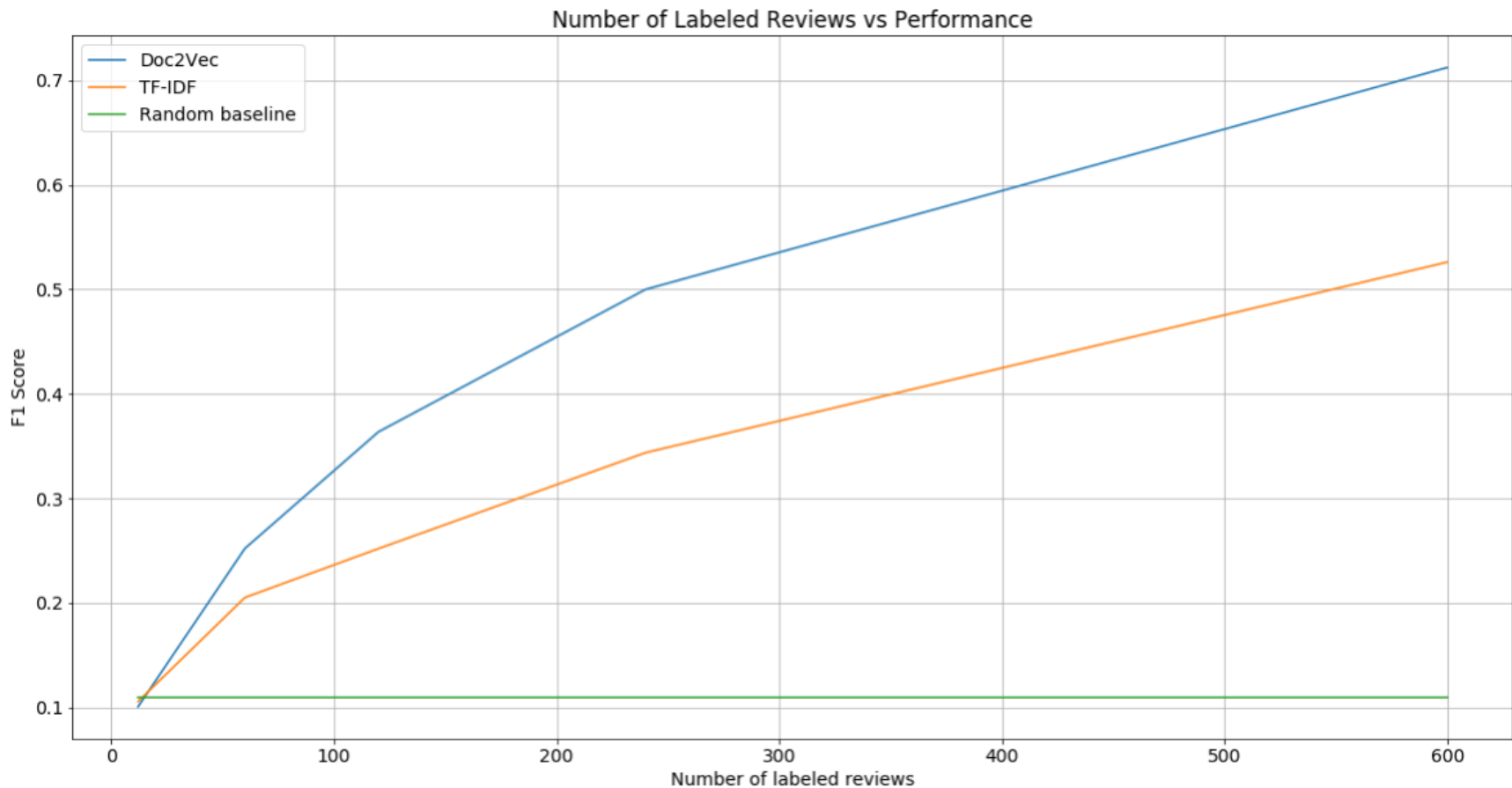
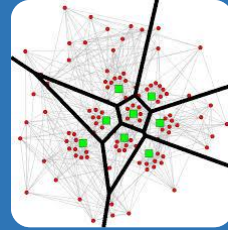
Threshold optimization



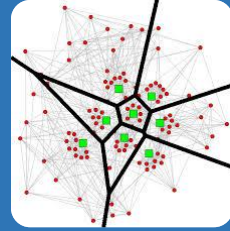
- The experiments are performed on Turkish restaurant reviews dataset.
- 50 labeled reviews are used as seeds for PPR for each label
 - Totally 600,
 - 50% of the dataset
- According to the results:
Best threshold for both TF-IDF and Doc2Vec is 0.003



Experimental Results



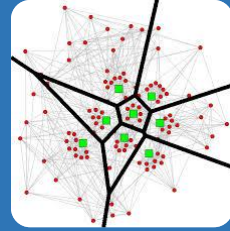
Experimental Results



Method	# of labeled reviews	Average F1 Score
Random baseline	0	0.11
LSTM + Logistic Regression[3]	1.1K (90%)	0.61
TF-IDF + PPR	12 (1%)	0.1
	60 (5%)	0.2
	240 (20%)	0.34
	600 (50%)	0.52
Doc2Vec + PPR	12 (1%)	0.1
	120 (10%)	0.25
	240 (20%)	0.5
	600 (50%)	0.71

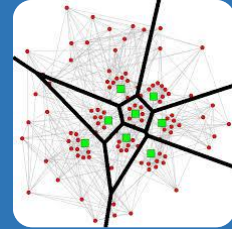
[3] Tamchyna, Aleš, and Kateřina Veselovská. "Ufal at semeval-2016 task 5: recurrent neural networks for sentence classification." *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 201

Conclusion



- With proposed method (Doc2Vec similarity graph + PPR)
 - High accuracies are possible with smaller number of labeled reviews
- Nodes without connection decreases the performance:
 - 226 unconnected nodes in Doc2Vec similarity graph leads to low recall values
 - Average Precision: 0.98,
 - Average Recall: 0.64,
 - Average F1: 0.71
- The proposed system is easily extendible. Experiments should be performed on other languages and domains

References



- [1] Pontiki, Maria, et al. "SemEval-2016 task 5: Aspect based sentiment analysis." *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 2016.
- [2] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International Conference on Machine Learning*. 2014.
- [3] Tamchyna, Aleš, and Kateřina Veselovská. "Ufal at semeval-2016 task 5: recurrent neural networks for sentence classification." *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016.
- [4] Shayan A Tabrizi, Azadeh Shakery, Masoud Asadpour, Maziar Abbasi, and Mohammad Ali Tavallaie. "Personalized pagerank clustering: A graph clustering algorithm based on random walks." *Physica A: Statistical Mechanics and its Applications*, 392(22):5772–5785, 2013.