sentence/query similarity deneysel sonuçlar

* Bu çalışmada sentence/query similarity ölçümünde 4 farklı yöntemden elde edilen sonuçlar karşılaştırılmıştır.

* Dataset: https://github.com/savasy/QuestionParaphrasesForTurkish/blob/master/TurkQP.csv Savaş Yıldırım, Turkish question paraphrase dataset

* 1. yöntem - kelime kök eşleşmesi jaccard puanı: Verilen q1 ve q2 querylerindeki unique kelimeler üzerinden jaccard puan hesaplanmıştır. (https://en.wikipedia.org/wiki/Jaccard_index)

* 2. yöntem – tf-idf vektör benzerliği: Verilen q1 ve q2 queryleri için tf-idf vektör elde edilip cosine similarity hesaplanmıştır.

* 3. yöntem – fasttext vektör benzerliği: Verilen q1 ve q2 queryleri için fasttext vektör elde edilip cosine similarity hesaplanmıştır.

* 4. yöntem – LSTM tabanlı Siamese Network: Verilen q1 ve q2 query temsilleri lstm siamese network ile eğitilmiş ve query similarity hesaplanmasında kullanılmıştır. Embedding layer fasttext kelime vektörleri ile ilklendirilmiştir.

* Yapılan denemeler sonucunda aynı test seti üzerinde 4. yöntemden diğerlerine göre daha yüksek başarım değeri (accuracy) elde edilmiştir. Detaylı sonuçlar tabloda verilmiştir.

* 5. yöntem olarak BERT çıktılarının Siamese networke verilmesi ve 4. yöntem ile sonuçların karşılaştırılması amaçlanmaktadır.

* Model kodlarına https://github.com/gurkan08/BertForSequenceClassification adresinden ulaşabilirsiniz.

Metod	Accuracy (test set)
1. yöntem (jaccard)	0.74
2. yöntem (tfidf)	0.72
3. yöntem (fasttext)	0.48
4. yöntem (lstm-siamese)	0.82

--
09/12/2020
Gürkan Şahin