Cmpe 493 Introduction to Information Retrieval, Spring 2018
Term Project - Turkish Sentiment Analysis for Twitter, Due: 22/05/2018 (Thursday), 23:59

In this assignment, you will implement a sentiment analysis system for Turkish tweets. We will share some libraries and papers for Turkish NLP and sentiment analysis with you. However, you are not restricted with these libraries and you are not obligated to use them. You **cannot** submit any available sentiment analysis codes or tools as your base code, but you can use them to compare your results at the presentation.

Submission: You should submit your codes through email. You should send an email with the title "CmpE 493 Term Project" until the deadline to abdullatifkoksal@gmail.com (No late submission allowed).

Demo: There will be a demo session on May 24 between 10:00 and 17:00. It will be held in AILab at the CmpE building. You should come with your computer with the libraries, you used in your code, installed. You will be given a test dataset and we are expecting labels as output. Please make sure that the submitted code by email is running on your computer and generating outputs in time as we need to get your output file until 17:00.

Format of the Test Data: The tweets will be separated by a new line in the test file. Your code should produce an output file which includes the labels of tweets in each line. The labels for positive sentiment is 1, neutral sentiment is 0, and negative sentiment is -1.

Input:
tweet-1
tweet-2
...
tweet-n

Output:
label-tweet-1
label-tweet-2
...
label-tweet-n

Presentation: You must prepare and give a presentation about your sentiment analysis model on the final exam date/time (May 25th, Slot 1, BM A3). Each group will give a 6 minutes presentations and 2 minutes Q/A part. You should send your presentation in pdf format to abdullatifkoksal@gmail.com until **May 24th, 23:59** with the title "CmpE 493 Final Presentation *GroupID*". You should describe your model and features in the presentation. You should also discuss your observations on the dataset. Also, you can include sections as what haven't you done and what could you have done for better performance. You need to comment on your results, too.

Sources:
Like It or Not: A Survey of Twitter Sentiment Analysis Methods
https://dl.acm.org/citation.cfm?id=2938640
BOUNCE: Sentiment Classification in Twitter using Rich Feature Sets
https://www.cmpe.boun.edu.tr/~ozgur/papers/S13-2093.pdf
Turkish NLP Datasets by Yildiz Technical University
http://www.kemik.yildiz.edu.tr/?id=28
Turkish NLP Resources
http://www.denizyuret.com/2006/11/turkish-resources.html
Turkish Positive and Negative Words
https://moodle.boun.edu.tr/pluginfile.php/301953/mod_resource/content/2/positiveNew.txt
https://moodle.boun.edu.tr/pluginfile.php/301954/mod_resource/content/1/negativeNew.txt

If you want to implement spelling correction you can follow this tutorial.
https://norvig.com/spell-correct.html
You can use Zargan Lexical Database to feed spelling correction with Turkish words.
http://st2.zargan.com/duyuru/Zargan_Linguistic_Resources_for_Turkish.html
Zemberek library includes several NLP algorithms for the Turkish language for Java
https://github.com/ahmetaa/zemberek-nlp
Fasttext pretrained word vectors trained on Turkish Wikipedia
https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
Word2vec pretrained model trained on Turkish Wikipedia. You can train your model as well.
https://github.com/akoksal/Turkish-Word2Vec