Homework 3 – Analysis

Author: Jianxiang Fan ***Email: jianxiang.fan@colorado.edu***

■ *Cross Validation*

Randomly pick **25% data** from training set as development test set. Run **100 times**. Compute **(mean accuracy, standard deviation)**

■ *Features of the text*

a) 1-gram (Original baseline)

--feat=word (0.65304924242424245, 0.010191640598415616)
By 1-gram, some *spoiler words* are picked out, such as 'die', 'kill', 'shoots', 'reveals'...

b) 1,2-grams and 1,2,3-grams

--feat=word --ngmax=2: (0.67034361471861492, 0.0080401400022678173)
--feat=word --ngmax=3: (0.67305465367965367, 0.0074567310637176989)
By multi-grams, some *spoiler phrases* are picked out, such as 'as result', 'in episode', 'then the'...

c) Tf-idf

--feat=word --tfidf: (0.68758387445887437, 0.0065734721566692066)
--feat=word --tfidf --ngmax=2: (0.68583603896103895, 0.0075764701852537161)
--feat=word --tfidf --ngmax=3: (0.680386904761905, 0.0065812184630768189)
Adjust for the fact that some features appear more frequently in general.

d) Stop-words

--feat=word --stop --tfidf: (0.68089826839826872, 0.0058382656303679076)
--feat=word --stop --tfidf --ngmax=2: (0.68030573593073596, 0.0073834488127865298)
Doesn't work well with tfidf and n-gram

e) Use stem instead of word

--feat=stem --tfidf: (0.68565205627705605, 0.0076519107484513803)
--feat=stem --tfidf --ngmax=2: (0.68632575757575753, 0.0071764162294187953)
Doesn't make too much difference from using words.

f) 1st and 3rd person words

--feat=1st:3rd --tfidf: (0.56965638528138529, 0.028768734047437843)
--feat=1st:3rd:word --tfidf --ngmax=2: (0.68121482683982659, 0.008637510824300778)
3rd person words tend to appear in spoiler, 1st person words not , while doesn't make too much difference in whole.

g) Sentence Length

--feat=length --tfidf: (0.53001082251082254, 0.01406737416544027)
--feat=length:word --tfidf --ngmax=2: (0.68291937229437205, 0.0072445428894296971)
It looks like that sentence length is not quite related.

■ *Features of the movie/series (meta info from imdb)*

Data provided by omdbapi (http://www.omdbapi.com/): Title, Year, Rating, Released, Runtime, Genre, Director, Writer, Actors, Language, Country, Rating, Votes

h) Genre (Very useful feature)

--feat=genre --tfidf: (0.6468614718614718, 0.013079464400758183)
--feat=genre:word --ngmax=2 --tfidf: (0.71104978354978354, 0.0060908385172686761)
Pos: Drama, Crime, Mystery, Horror, Sci-Fi, Thriller, Romance,...
Neg: Comedy, Family, Game-Show, Reality-TV, Documentary, Talk-Show, Music,...

i) Rating and votes

--feat=rating:votes --tfidf: (0.59491612554112561, 0.013039985703541737)
--feat=rating:votes:word --ngmax=2 --tfidf: (0.69738906926406941, 0.006953660595678892))
Shows with high rating and votes may contain more spoiler.
Pos: ~V:11,~Ra:8,~V:10,~V:5,~Ra:9,... **Neg: ~Ra:4,~Ra:3,~V:7,~V:4,~V:3,...**

j) Year

--feat=year --tfidf: (0.59199945887445882, 0.033151292474392025)
--feat=year:word --ngmax=2 --tfidf: (0.69626082251082233, 0.0065616665345398523)
More recent shows tend to contain more spoiler (20-year as a class)
Pos: ~Y:2000,~Y:1940,~Y:1980,... **Neg: ~Y:1900,~Y:1960,~Y:1920,...**

k) Put all together

--feat=word:genre:rating:votes:year:runtime:country:lang --tfidf: (0.72684794372294381, 0.0058806499194479847)
--feat=word:genre:rating:votes:year:runtime:country:lang --tfidf --ngmax=2: (0.72452651515151534, 0.0056196838380655929)