Exercise 04

Lexicon based sentiment analysis with MapReduce

von Iuliia Nigmatulina, Yeon Joo Oh, Iliyana Kamenova

1) As documents we used 10 restaurants reviews, which were saved as separate files in one directory. We have independently annotated reviews as negative, positive, and neutral and calculated inter-rater agreement using Fleiss κ. In our case, Fleiss κappa was 0.7848, which means high agreement.

The script *SentimentAnal_ex1.py* defines sentiment score for each of the documents. The score is calculated according to the number of positive (+1) and negative words (-1) in a document, as well as the number of negation words (such as *no, not;* -1).

Some scores coincide very well with manual annotation, if following the principle that negative scores are for negative sentiment and positive are for positive (see appendix: e.g. docs 3, 4, 5, 6, 9). Scores for other docs are more ambiguous; it means that some calibration is needed to set the threshold that would fit the data in the best way. We decided that for our sample: pos >= 7; 3 <= neut <= 6; neg <= 2. With this calibration, scores of 7 of 10 docs coincide with annotators' decision, 1 negative document is defined as 'neutral' and 2 neutral documents are defined as 'negative'. Therefore, there is no positive-negative confusions, only negative-neutral and the boundary between neutral and negative categories is more subtle.

2) Scripts *execute_MapReducer.py* and *MapReducer.py* run the Map-Reduce version of the native lexicon-based sentiment analysis from the task 1. The script *visualise_SA.py* visualises the results of sentiment analysis. We also generated word clouds out of positive and negative words based on sentiment analysis.

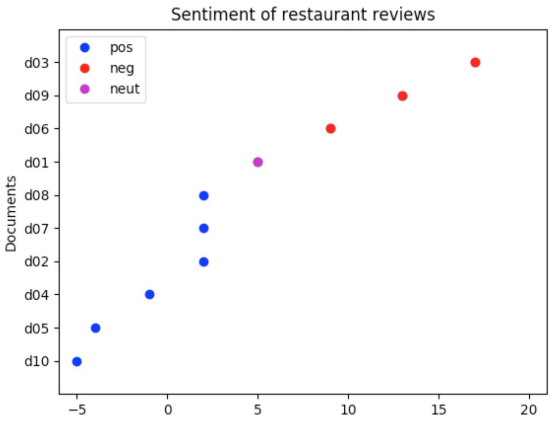

Figure 1: Map-Reduce plot

Figure 2: word cloud

APPENDIX

1) Fleiss kappa calculation

Annotation:	**d1**	**d2**	**d3**	**d4**	**d5**	**d6**	**d7**	**d8**	**d9**	**d10**
Julia:	neg	neg	pos	neg	neg	pos	neut	neg	pos	neut
Iliyana:	neg	neg	pos	neut	neg	pos	neut	neg	pos	neut
Yeon Joo:	neg	neg	pos	neg	neg	pos	neut	neg	pos	neg
Score:	5	2	17	-1	-4	9	2	2	13	-5

Inter-annotator agreement
Fleiss Kappa (k) statistics for nominal scales:

n = 3 annotators
k = 3 categories
10 data points (=documents)
N = 3 x 10 = 30
*(sum(pj^2) = 8x9 + 2x4 + 1x2 = 72 + 8 + 2 = 82)
P = (sum(pj^2) — N) / (N * (n — 1)) = (82 — 30) / (30 * 2) = 52 / 60 = 0.8666
Pe = sum(pi^2) = 0.09 + 0.25 + 0.04 = 0.38
pi = sum(p_column_i) / N

kappa = (P — Pe) / (1 — Pe) =
= (0.8666 — 0.38) / (1 — 0.38) =
= 0.4866 / 0.62 =
= **0.7848**

DOC	POS	NEG	NEUT	P
1	0	3	0	**9**
2	0	3	0	**9**
3	3	0	0	**9**
4	**0**	**2**	**1**	**5**
5	0	3	0	**9**
6	3	0	0	**9**
7	0	0	3	**9**
8	0	3	0	**9**
9	3	0	0	**9**
10	0	1	2	**5**
Total				**P = 82**
pi	**0.3**	**0.5**	**0.2**	
pi^2	**0.09**	**0.25**	**0.04**	**Pe = 0.38**