5.1.1:
Used 50/50 split in trainingdata as development set.
See Table 1 for full results.
Accuracy is 0.81, probably low due to small training sample. Almost every step after this increased the score, and enlarging the training set to 0.75 also helped in this.

5.1.2:
Affecting C seems not to influence the score positively, whether it is increased up till 10, or decreased down till 0.1
C is used for maximization of the margin, making for a broader terrain (the lower C is) around the axis that is not being taken into account.
The larger the C, the larger the penalty for errors because if a sample is found in the narrow margin, it is an indicator that the axis is not yet optimal.

5.1.3:
RBF scores at least as well as linear, and with altered values does it even better.
(acc. =0.81 with c=1 and gamma =0.7)
It certainly is not significantly outperformed by the linear kernel.

5.1.4:
Best script uses bigrams and english stopwords build into sklearn. Of all stemmers the Wordnet Lemmatizer performs best.
In the results it looks like it exchanges just 0.1 with the lancaster stemmer regarding negative precision versus positive recall,
but in the raw accuracy number it goes up about 0.3.
POS tagging took an incredible amount of time resources and is therefore not shown in the results.

5.2.1:
1: The CM (see CM_KM_ML.png) is distributed pretty well, except music that is almost always confused with health and dvd.
2: Health itself is confused with almost every other class, being the most confused class.
3: The scores obtained do not reach over 30%, so the system is not very good at clustering into these 6 classes (see table 2). The rand index is even lower than the V measure, not going higher than 18%. Using a stemmer or lemmatizer offers a slight improvement, but still the clusters are formed rather poorly. The amount of new centroid initiations also has a decreasing effect on the score, so somehow the first choice for the centroids is immediately the best. This can be explained by the compute parameter set to true, in which the system takes its time (and memory) to try to go for a good spot for the centroid placements. One very strange observation is that ngrams don't work at all, and limit the classifier in its performance a lot. Using bigrams, all instances are placed in one cluster, for a reason that is unclear till now.

5.2.2:
The articles about music and books will be clustered, but somehow kmeans clusters almost every instance into one cluster. The ri and v measure are extremely small, and do not go up by chainging any parameter or stemmer. Using sentiment it performs slightly better, but still the scores are rounded to zero (3 decimals).
the numbers are too small to make any significant difference visible, so a discriminative set of features to favour on over the other also is not possible.

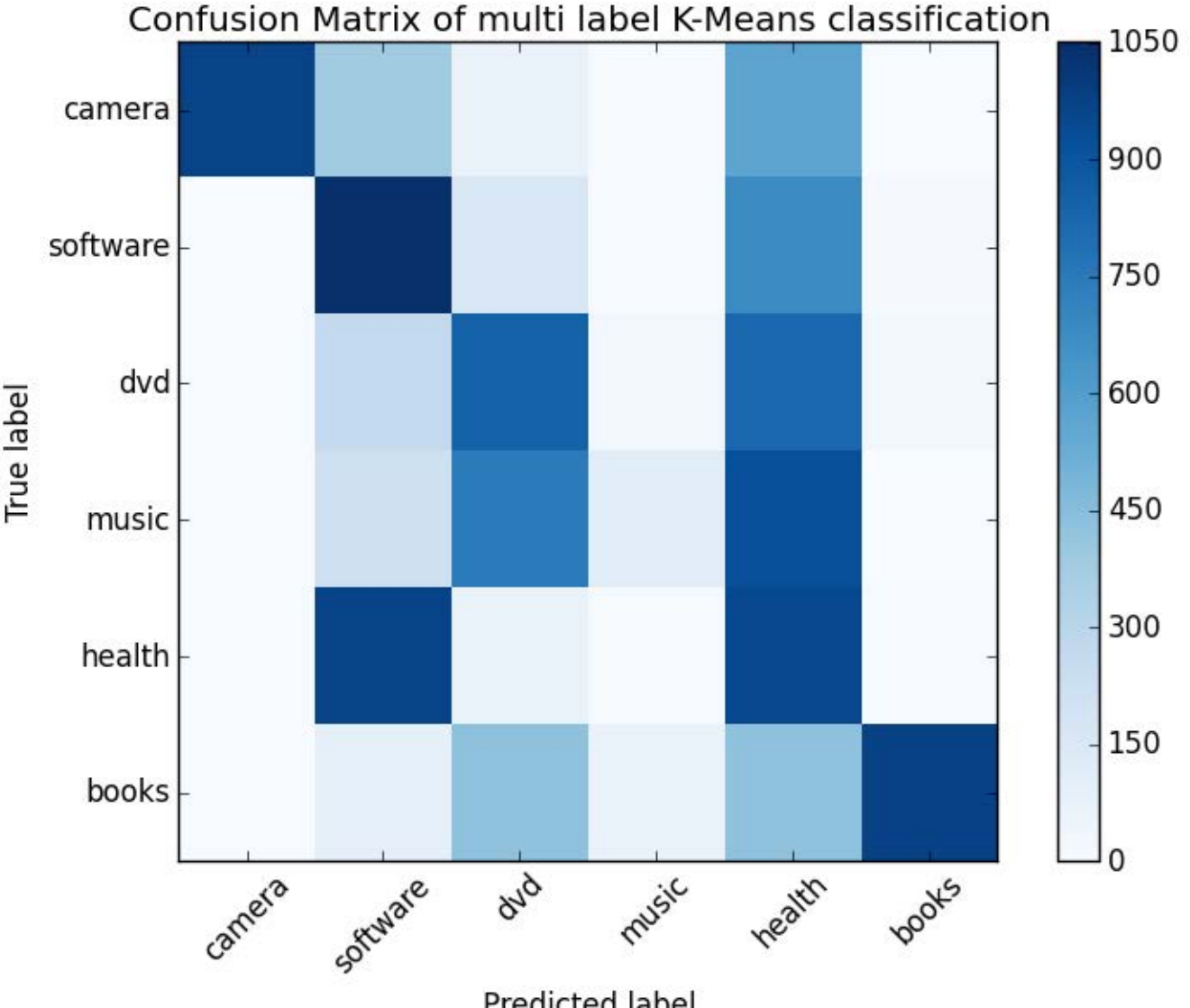
Table 1: SVM Results

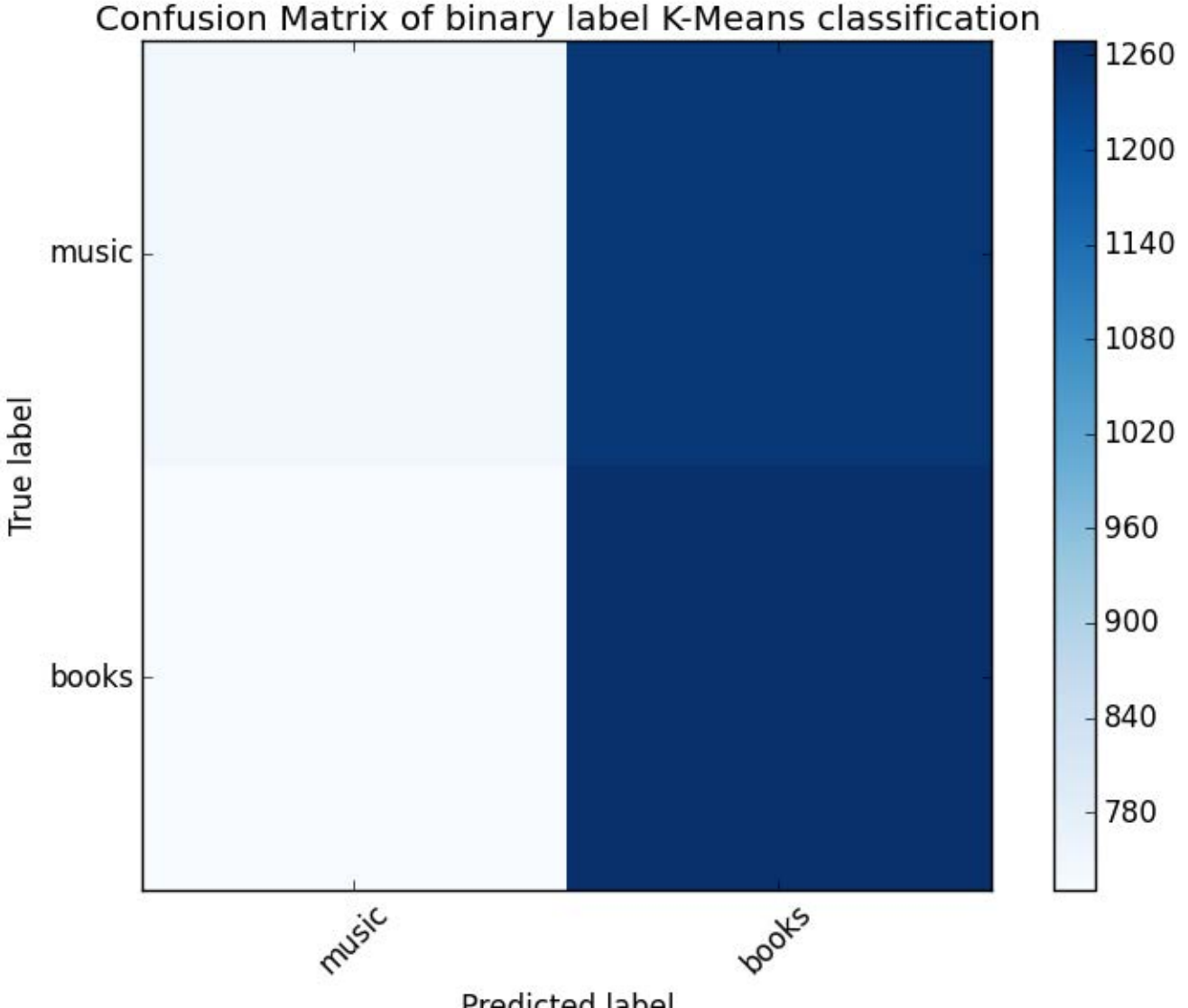
SVM 50/50 traindata	acc	prec	rec	f1
Linear, c=1.0	0.81			
POS		0.80	0.81	0.81
NEG		0.81	0.81	0.81
Linear, c=0.5	0.81			
POS		0.81	0.81	0.81
NEG		0.81	0.81	0.81
Linear, c=0.1	0.76			
POS		0.76	0.77	0.76
NEG		0.76	0.75	0.76
Linear, c=2.0	0.80			
POS		0.81	0.80	0.80
NEG		0.80	0.81	0.80
Linear, c=5.0	0.79			
POS		0.80	0.78	0.79
NEG		0.78	0.80	0.79
Linear, c=10.0	0.79			
POS		0.80	0.79	0.79
NEG		0.78	0.80	0.79
Rbf,g=0.7, c=1.0	0.81			
POS		0.82	0.81	0.81
NEG		0.81	0.81	0.81
Rbf,g=0.7, c=0.8	0.80			
POS		0.81	0.80	0.80
NEG		0.80	0.81	0.80
Rbf,g=0.5, c=1.0	0.81			
POS		0.81	0.81	0.81
NEG		0.81	0.81	0.81
Rbf,g=0.1, c=1.0	0.771			
POS		0.78	0.76	0.77
NEG		0.76	0.79	0.77
sigmoid, g=0.8, c=1 +bigrams	0.82			
POS		0.81	0.83	0.82
NEG		0.82	0.81	0.81
OptSVC= previous w/0,75 train	0.83			
POS		0.81	0.86	0.83
NEG		0.86	0.80	0.83
OptSVC + Porterstemmer	0.83			
POS		0.80	0.87	0.83
NEG		0.86	0.80	0.83
OptSVC + Snowball(stopF)	0.83			
POS		0.80	0.87	0.83
NEG		0.86	0.80	0.83
OptSVC + lancaster	0.84			
POS		0.81	0.87	0.84
NEG		0.87	0.81	0.84
OptSVC + lemmatizer	0.84			
POS		0.81	0.88	0.84
NEG		0.88	0.80	0.84

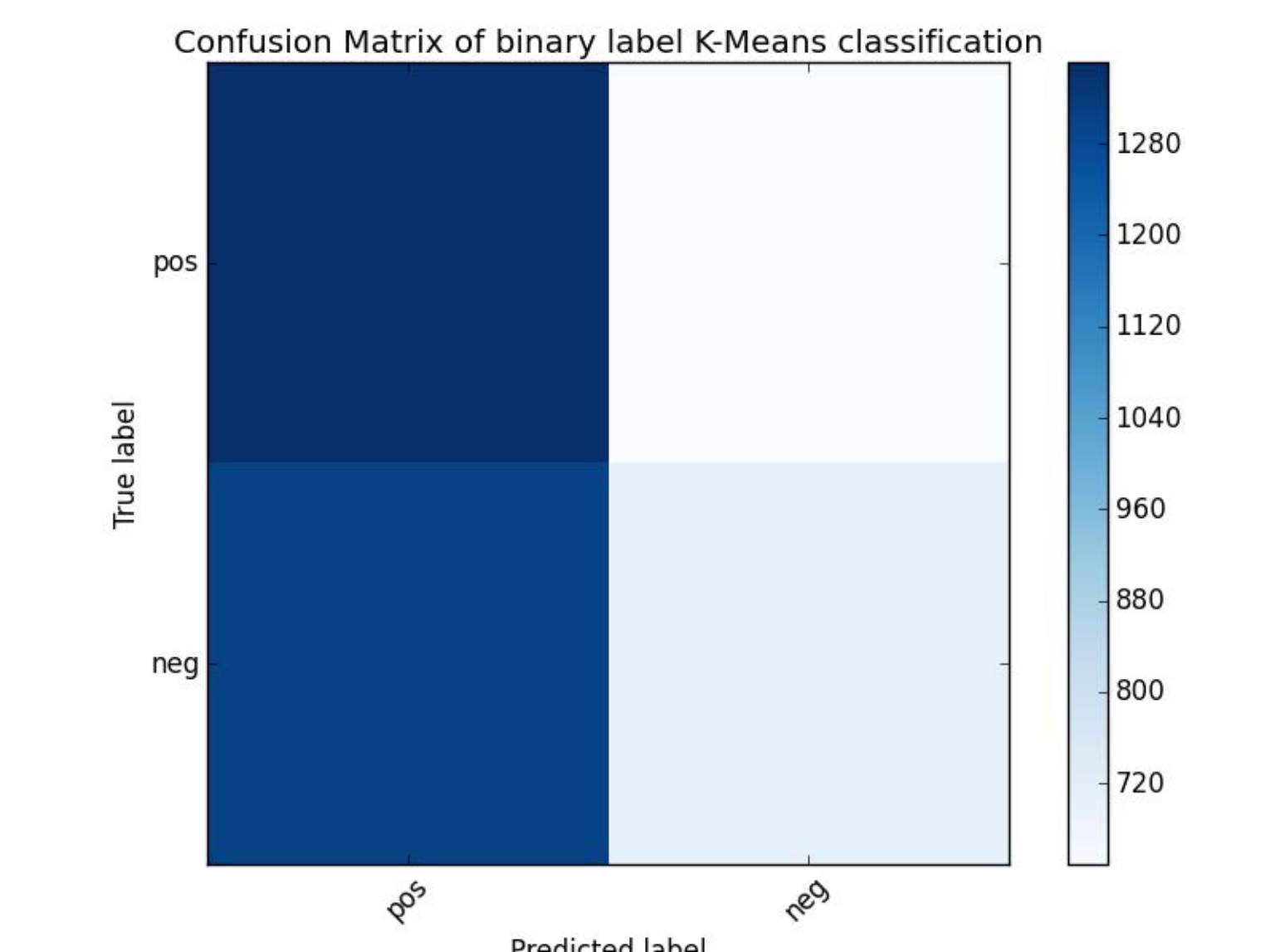
Table 2: KM results

Classifier settings (Multi)	RI	V
init=1, verbose=1,precdist (=optKM)	0,135	0,241
init=100, verbose=1	0,105	0,183
init=10, verbose=1	0,091	0,170
optKM+WNlemmatizer	0,051	0,143
optKM+lancasterstemmer	0,124	0,211
optKM+Porterstemmer	0,175	0,266
optKM+Snowballstemmer	0,119	0,233
optKM+Snowball(stop=True)	0,160	0,280

Classifier settings (Binary)	RI	V
Best from multi (class labels)	0,000	3,15E-06
Best from multi (sentiment labels)	0,000	0,000
optKM sentiment	0,000	0,002


Confusion Matrix of multi label K-Means classification
camera
software
dvd
music
health
books
True label
camera
software
dvd
music
health
books
Predicted label
1050
900
750
600
450
300
150
0


Confusion Matrix of binary label K-Means classification
True label
music
books
music
books
Predicted label
1260
1200
1140
1080
1020
960
900
840
780


Confusion Matrix of binary label K-Means classification
True label
pos
neg
pos
neg
Predicted label
1280
1200
1120
1040
960
880
800
720