Web Intelligence - Social Media Miniproject

Erik Sidelmann Jensen
ejens11@student.aau.dk

Lasse Vang Gravesen
lgrave11@student.aau.dk

Dennis Jakobsen
djakob11@student.aau.dk

1 Social Media Miniproject

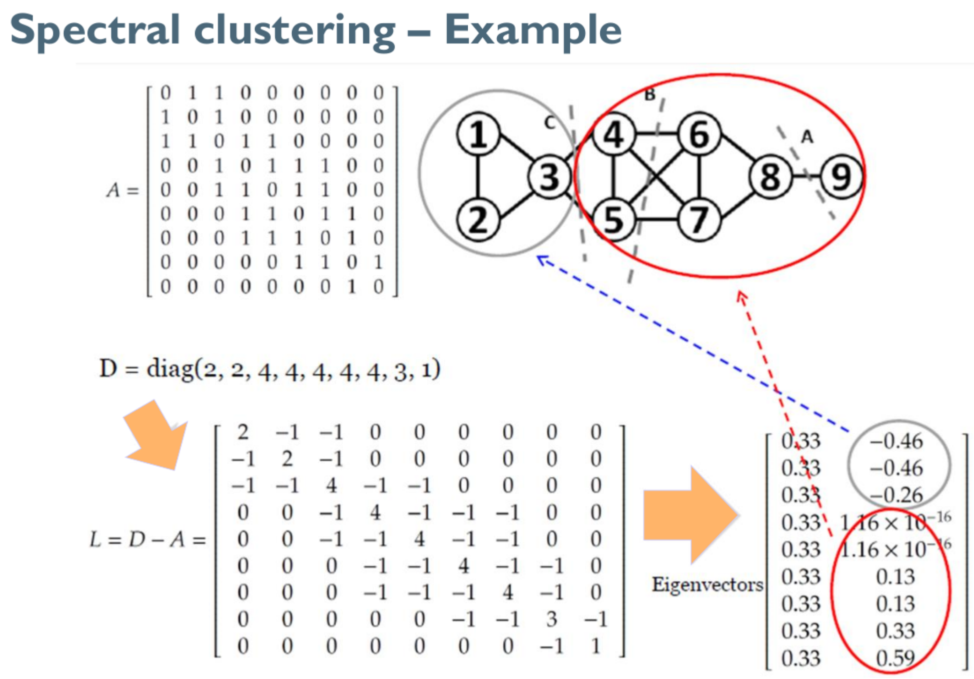
1.1 Network Communities Partitioning

There are several approaches to discover communities. For example Simple local(member) based clustering, Global(group) based clustering, and Hierarchical clustering. We did community partitioning using spectral clustering, which is community detection by graph cutting. The advantages of using spectral clustering is that it is easy to implement, gives good results, and is reasonably fast for sparse data sets of 1000+ elements. The disadvantages of using this approach is that it is computationally expensive for large data sets.

The general algorithm behind spectral clustering is this:

1. Given an adjacency matrix, A, representing a graph and an diagonal matrix D where the diagonal entry i is the sum of i'th row in A. Compute the unnormalized Laplacian L using $L = D - A$.
2. Compute the second eigenvector v_2 of L, the one with the second smallest eigenvalue - not 0.
3. The j'th value in the eigenvector corresponds to the j'th node.
4. Order the nodes according to their eigenvector values.
5. Cut at the largest gap.

An example of this can be seen in the figure below.



Spectral clustering – Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D = \mathrm{diag}(2, 2, 4, 4, 4, 4, 4, 3, 1)$$

$$L = D - A = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Eigenvectors

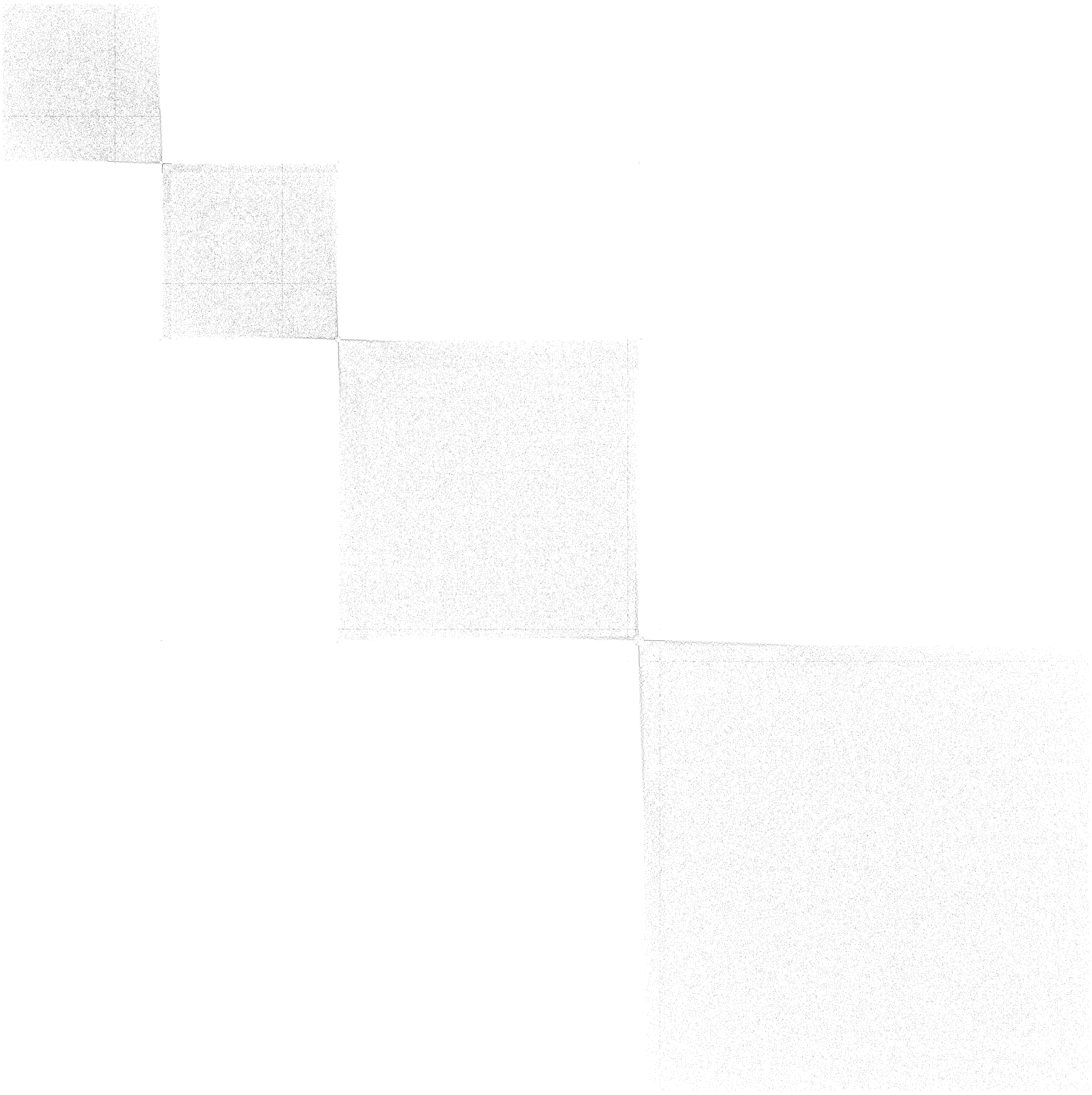
$$\begin{bmatrix} 0.33 & -0.46 \\ 0.33 & -0.46 \\ 0.33 & -0.26 \\ 0.33 & 1.16 \times 10^{-16} \\ 0.33 & 1.16 \times 10^{-16} \\ 0.33 & 0.13 \\ 0.33 & 0.13 \\ 0.33 & 0.33 \\ 0.33 & 0.59 \end{bmatrix}$$

To do this in practice the following was done:

A User class was created to hold the basic information, the Username, the list of friends, the review, the summary, and a sortable eigen value. Using this as a model the file containing the friendships and reviews could be loaded into a list of User objects. From this list of Users, the adjacency matrix A is created based on who they are friends with. From A, D is created The laplacian is calculated using L = D - A. The eigen decomposition is then found on L, and the second smallest eigenvector is found. This is the computation of the second vector. The corresponding eigenvector-value is then assigned to a User object in the user list, after which the user list can be sorted. Then the largest gap can be cut.

To find more communities, this is then run recursively until a threshold gap of 0.7 is reached after which it stops and returns the list of communities. Alternatively to running it recursively, multiple eigenvectors could have been used to perform the cut. Also alternatively to the threshold gap, modularity could have been used.

The results of this is 4 very segregated communities, on the first run-through the communities are obvious as can be seen in the figure below. We do not know if there are any flaws in our implementation but it seems to work.



The spectral clustering algorithm was used because it is simple to implement, and works well.

1.2 Sentiment Analysis

Sentiment analysis of text was done using a Naive Bayes Classifier. We could also have used a classifier called 'Conditional Maximum Entropy', but for our purposes Naive Bayes was sufficient.

This was the procedure: First step is tokenising the test data and constructing features, and for sentiment analysis it is important to remember things like emoticons and negative words because those have an impact on the rest of the content. As such we took a sentiment sensitive tokeniser made in Python and converted it to C#, and added sensitivity to negative words, in that every token followed by a negative word would get "_NEG" appended until the end of the sentence(punctuation or an emoticon).

The second step was parsing the learning data, which was done line by line, skipping the identifying text and taking the rest after having trimmed the result. Tokenising is also done in the parser, the tokeniser is called and these tokens are saved in a properties. Finally the known score of the review block is determined by the score. If the score is 3, the review is discarded. These reviews are then split into N partitions for easy learning and testing.

The third step is then constructing the Naive Bayes Classifier, which is essentially just counting. Specifically these properties of the learning data is determined through counting:

Estimation of the probability for all classes: $p(c)$. This is done through counting the amount of reviews available: N. The number of reviews with sentiment c: $N(c)$. We also have $|C|$, the amount of classes. Finally $p(c)$ can be estimated through

$$p(c) = \frac{N(c) + 1}{N + |C|}$$

We also estimate probability for all words in corpus X and all possible sentiment classes in C: $p(x_i|c)$. To do this we need to count the number of times x_i appears across all reviews with sentiment c: $N(x_i, c)$. We also need $|X|$, the size of the vocabulary. Using this we can then calculate $p(x_i|c)$:

$$(x_i|c) = \frac{N(x_i, c) + 1}{N(c) + |X|}$$

We also use Laplace smoothing, which ensures that no probability is exactly zero.

Then the score for a review can be determined through two tricks: First calculate score for the "empty" review because most words from the vocabulary are not present in a given review(Trick 1). This is done using summation instead of product to avoid numerical instability(Trick 2).

$$s^*(empty, c) = (\sum_{i=1}^{n} p(\text{not } x_i|c)) + p(c)$$

The score can then be calculated using:

$$s(x, c) = s^*(empty, c) + (\sum_{j=1}^{k} \frac{p(x_j|c)}{p(\text{not } x_j|c)})$$

And that is how the Naive Bayes Classifier works, you first use the learning data to construct the probabilities and then you use those probabilities to determine the score for individual reviews.

To test the classifier we used all possible combinations of test and learning data, where test data was 1 partition and learning data was the rest. So over N partitions, that means we iteratively ran over the partitions, selected the first for testdata, the rest for learning data and then in the second iteration we select the second for testdata and so on. This is also called cross validation.

Using this to improve the feature construction, we ended up with an accuracy rate of roughly 91% and an error rate of roughly 9% for each iteration. Though the classifier ended up being biased towards positive reviews in that it correctly classified them roughly 94% of the time, and for negative reviews it classified them correctly roughly 77% of the time.

Alternatively to the Naive Bayes Classifier, MaxEnt could have been used and would probably have been a good choice since it apparently gives better results than Naive Bayes. It is more complicated however, and Naive Bayes works fine for our purposes.

1.3 Evaluation of Reviews

We then used the community partitions and the sentiment classifier to evaluate the reviews.

We did this by using the entirety of the learning data to teach the Naive Bayes Classifier. We then ran through the list of users and classified the reviews of users that had reviews/summaries using the Naive Bayes Classifier. Then we determine the communities using the spectral clustering. Then we ran through the list of users again, this time only focusing on the users that did not have any reviews or summaries, and based on their friends average review scores we determine if they are likely to purchase. A user named 'kyle' and friends in other communities count 10 times more than a regular friend, finally we determine if they are likely to purchase by checking if the average is more than or equal to 3, and if it is the user is likely to purchase otherwise not.

Our result in summary is then as follows: We classified 1347 user reviews out of 1644 as 'Positive', and 297 user reviews as 'Negative'. Of the users that did not have a review or a summary, we classified 2451 out of 2575 as likely to purchase and 124 as unlikely to purchase.