Inference of User Interests from Multiple Social Networks

...

Steven Kester Yuwono (A0080415N)
Choo Jia Le (A0116673A)

Methodology

Twitter: JSON format

No	Features	Descriptions
1	**Text**	**The content of the tweet**
2	Retweets	The number of retweet for this particular tweet
3	Timezone	The timezone of the user's location when this tweet is posted
4	Friends Count	The number of users this account is following (AKA their "followings")
5	Followers Count	The number of followers this account currently has.
6	Favourites Count	The number of tweets this user has favorited in the account's lifetime.
7	User ID	ID of the person posting this tweet

Methodology

Facebook: Plain Text, contains many unnecessary words

Timothy Nichols changed his profile picture. Share 9 people like this. Remove Lisa Nichols Well how did ya like being fired? Hahaha 1 July 25 at 2:21am Remove Timothy Nichols Haha. Well said July 25 at 2:32am Remove Trudy Patterson Boy I wish I would have done that a long time ago. 1 July 25 at 9: 28pm"

After removal using regular expressions

Lisa Nichols Well how did ya like being fired? Hahaha Timothy Nichols Haha. Well said Trudy Patterson Boy I wish I would have done that a long time ago.

Methodology

LinkedIn: HTML Format. Detect each section of the profile.

No	Section Title	Descriptions
1	Full Name	Full name of the user
2	Location	Geographical location
3	Current Position	The position currently held
4	**Summary**	**The summary paragraph/text**
5	**Skills**	**List of skills acquired**
6	Current Experience	Current job roles and description
7	Past Experiences	A list of past job roles and descriptions
8	Honors and Awards	List of honors and awards attained
9	Organizations	List of organizations which the user belongs to
10	**Interests**	**List of interests**

Methodology

Preprocessing of Data:

1. Tokenization
2. Convert all words to lowercase
3. Removed url links with regex
4. Removed stopwords
5. Topic Modelling using LDA (Latent Dirichlet Allocation)
    a. MALLET toolkit

Methodology

- Early fusion
- Late fusion + Genetic Algorithm
  - Assigned weight to the confidence score with respect to each class in each source
  - The weight is learned by using **Genetic Algorithm**.

Methodology

- Learning the model using SVM
- There is an unbalanced dataset problem.
  - 20% of negative training dataset and 80% positive training dataset,
  - the classifier's positive-class prediction value will be very close to 80%.
- Hence uses minority-oversampling technique

Evaluation

- **P@K:** the proportion of the top K recommended interests that are correct
- **S@K:** the mean probability that a correct interest is captured within the top K recommended interests.
- **R@K:** the proportion correct number of predicted interests (in the top K) out of the actual number of interests for that particular user

Results

Number of Topics	5	10	20	35	50	100	200	300	1000
Facebook P@5	0.356	0.367	0.339	0.360	0.343	**0.355**	0.332	0.348	0.333
Facebook R@5	0.257	0.265	0.245	0.260	0.248	**0.257**	02.40	0.252	0.241
Facebook S@5	0.947	0.907	0.887	0.887	0.867	**0.907**	0.893	0.887	0.927
LinkedIn P@5	0.251	0.244	0.392	0.492	0.489	**0.501**	0.476	0.497	0.489
LinkedIn R@5	0.181	0.176	0.284	0.356	0.354	**0.363**	0.344	0.360	0.354
LinkedIn S@5	0.693	0.713	0.940	0.953	0.933	**0.933**	0.940	0.967	0.960
Twitter P@5	0.392	0.449	0.440	0.423	0.409	**0.413**	0.415	0.403	0.420
Twitter R@5	0.284	0.325	0.318	0.306	0.296	**0.299**	0.300	0.291	0.304
Twitter S@5	0.913	0.927	0.940	0.940	0.913	**0.927**	0.913	0.893	0.860

Classifier's performance with respect to number of topics (LDA)

Results

Using Genetic Algorithm to train:

Facebook Weight	LinkedIn Weight	Twitter Weight	Late Fusion P@5	Late Fusion R@5	Late Fusion S@5
24.753	**695.971**	**141.733**	**0.476**	**0.344**	**0.947**
25.392	964.318	285.489	0.473	0.342	0.947
37.385	784.276	224.330	0.472	0.341	0.947
3.470	340.835	129.659	0.472	0.341	0.947

Late Fusion Best Weights

Result

K =	1	2	3	4	5	6	7	8	9	10	Ave
FB-P@K	0.307	0.300	0.318	0.330	0.355	0.368	0.354	0.358	0.360	0.350	0.340
LI-P@K	**0.440**	**0.483**	**0.513**	**0.500**	**0.501**	**0.493**	**0.485**	**0.467**	**0.456**	**0.431**	**0.477**
Tw-P@K	0.527	0.477	0.438	0.413	0.413	0.416	0.408	0.399	0.391	0.403	0.428
EF-P@K	0.433	0.417	0.364	0.360	0.375	0.397	0.405	0.401	0.401	0.395	0.395
LF-P@K	0.453	0.490	0.487	0.473	0.476	0.458	0.456	0.438	0.426	0.406	0.456

Overall P@K

Result

K =	1	2	3	4	5	6	7	8	9	10	Ave
FB-R@K	0.044	0.087	0.138	0.191	0.257	0.319	0.359	0.415	0.469	0.506	0.278
LI-R@K	**0.064**	**0.140**	**0.223**	**0.289**	**0.363**	**0.428**	**0.491**	**0.540**	**0.593**	**0.623**	**0.375**
Tw-R@K	0.076	0.138	0.190	0.239	0.299	0.361	0.413	0.462	0.509	0.582	0.327
EF-R@K	0.063	0.121	0.158	0.208	0.271	0.344	0.410	0.464	0.522	0.572	0.313
LF-R@K	0.066	0.142	0.211	0.274	0.344	0.397	0.462	0.506	0.554	0.587	0.354

Overall R@K

Result

K =	1	2	3	4	5	6	7	8	9	10	Ave
FB-S@K	0.307	0.540	0.687	0.813	0.907	0.940	0.973	0.980	0.993	1.000	0.814
LI-S@K	0.440	0.733	0.887	0.927	0.933	0.967	0.973	0.987	0.993	0.993	0.883
Tw-S@K	0.527	0.667	0.787	0.827	0.927	0967	0.980	0.993	1.000	1.000	0.867
EF-S@K	0.433	0.640	0.740	0.820	0.900	0.947	0.947	0.960	0.987	0.987	0.836
LF-S@K	**0.453**	**0.760**	**0.900**	**0.940**	**0.947**	**0.960**	**0.980**	**1.000**	**1.000**	**1.000**	**0.894**

Overall S@K

Results

K=	5	10
Bigram FB-P@K	0.368	0.353
Bigram LI-P@K	0.485	0.452
Bigram TW-P@K	0.412	0.413
Bigram EF-P@K	0.325	0.382
Bigram LF-P@K	0.465	0.406
Unigram FB-P@K	**0.355**	**0.350**
Unigram LI-P@K	**0.501**	**0.431**
Unigram TW-P@K	**0.413**	**0.403**
Unigram EF-P@K	**0.375**	**0.395**
Unigram LF-P@K	**0.476**	**0.406**

Bigram and Unigram Comparison

Conclusion

- **LinkedIn** is the most useful source, followed by Twitter, and then Facebook.
- **100** is the best number of topic for topic modelling in this case.
- Late-Fusion's performance is limited by the performance of the best source
  - Very difficult to outperform all the three sources significantly
- Noise (words that are not important) in the data causes a drop in the performance of the classifier
- Bigram was explored and found to be not very useful in this task
- LinkedIn performs the best overall, followed by Late-Fusion and, Twitter.