Latent Topics in Microposts

Andrés García-Silva, Victor Rodriguez-Doncel, Oscar Corcho: ***Semantic Characterization of Tweets Using Topic Models: A Use Case in the Entertainment Domain. International Journal on Semantic Web and Information Systems (IJSWIS) , 9 (3***)

Introduction to Topic Models

Document Collection (sentences) :

- I teach math and science — Topic 1 100 %
- I prefer science and literacy — Topic 1 100 %
- Spring and autumn are my favourite seasons — Topic 2 100 %
- The weather during winter and autumn is awful — Topic 2 100 %
- They learn about spring in science class — Topic 1 60 % — Topic 2 40 %

LDA will discover the topics in this sentences (or documents).
However you should define the number of topics.

For 2 topics LDA would produce something like this:

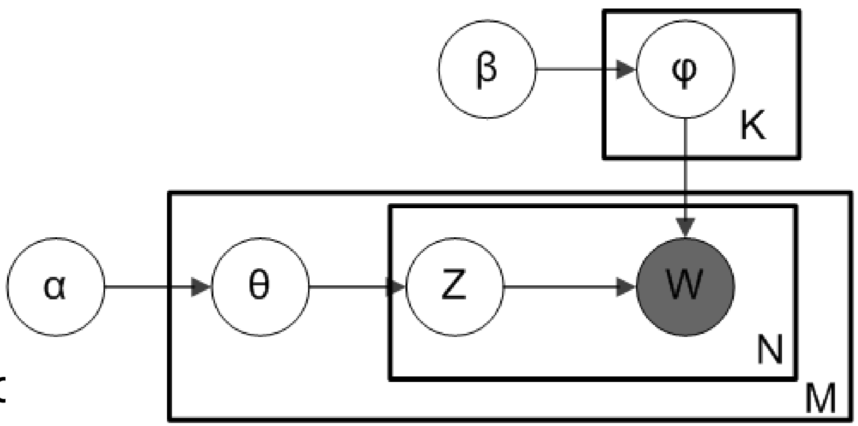
Topic 1={math 15%, science 30%, literacy 15%, teaching 10%, class 10%..}

Topic 2={spring 30%, autumn 30%, winter 10%, season 10% weather 10%..}

LDA – Latent Dirichlet Allocation

Allows sets of observations (words) to be explained by <u>unobserved</u> groups (Topics) that explain why some parts of the data are similar.

- M: Documents
- N: Words
- K: Topics
- <u>θ_i: topic distribution for document *i*</u>
- <u>Φ_k : is the word distribution for topic *k*</u>
- $z_{i,j}$: is the topics for the j-th word in the doc
- $w_{i,j}$: is the j-tj word in the doc I
- α: Dirichlet prior on the per-document topic distributions.
- β: Dirichlet prior on the per-topic word distribution



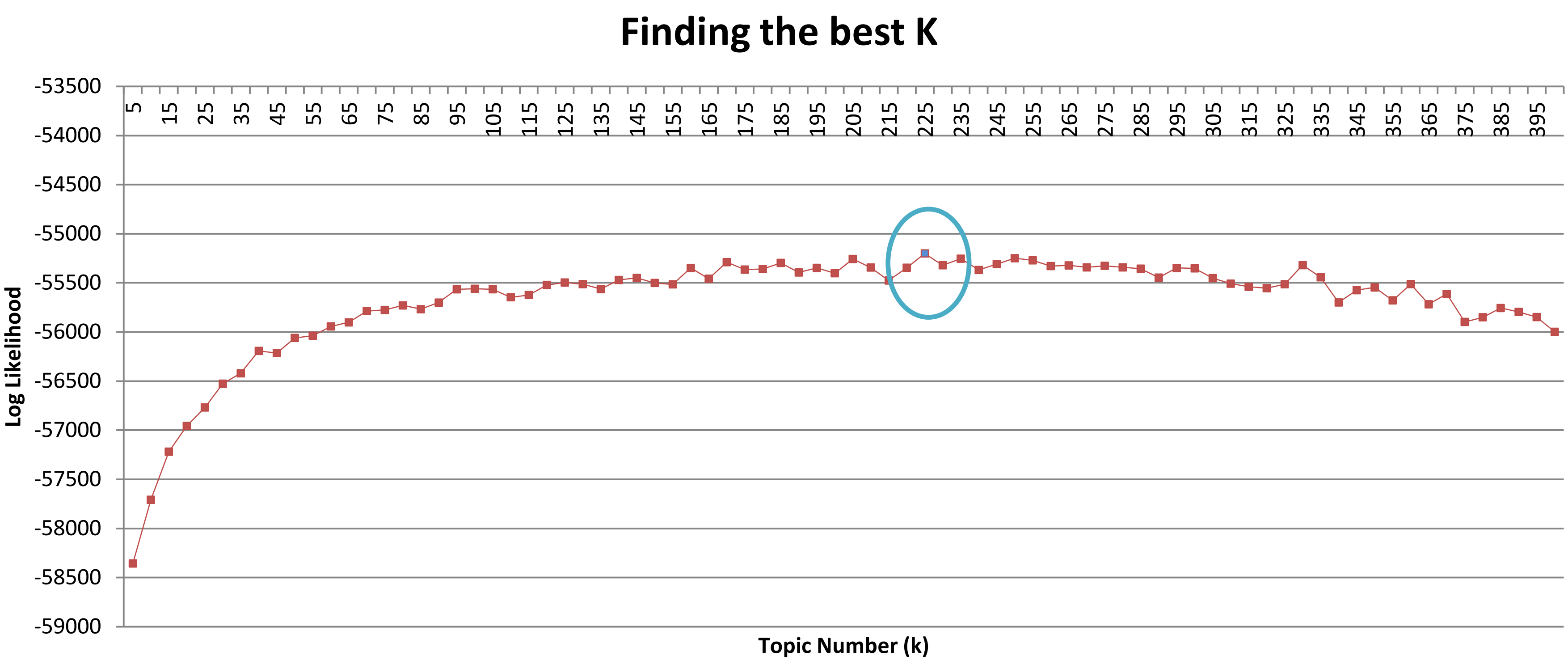
$$p(\theta, \phi, \mathbf{z} | \mathbf{w}, \alpha, \beta)$$

Bayesian Inference problem

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John. ed. "Latent Dirichlet allocation". *Journal of Machine Learning Research* **3** (4–5): *pp.* 993–1022.

Topic Models: Tweets

- 144 theatre plays announced in TimeOut London.
- We collected ~7000 geo-localized (London) tweets



Topic Models: Tweets

- K = 225

"Jamie Parker as Henry V at #TheGlobe was AMAZING. He said 'Cry God for Harry, England & St.Geooorge!' Then we won 6 Gold medals! #Olympics" ok

Topic 46, 0.379	Topic 51, 0.324	Topic 55, 0.0549
"v";0.202	"v";0.100	"amazing";0.065
"henry";0.195	"henry";0.094	":)";0.061
"globe";0.061	"england";0.040	"many";0.061
"parker";0.028	"bbc";0.033	"see";0.051
"jamie";0.025	"olympics";0.033	"seen";0.041
"watch";0.023	"crown";0.033	"times";0.041

Topic Models: Tweets

"Tonight I'm going to see a play I've never seen before... 'A Midsummer Night's Dream'."

Topic 94, 0.286	Topic 56, 0.197
"night";0.157	"see";0.115
"dream";0.153	"going";0.054
"midsummer";0.134	"tonight";0.051
"theatre";0.046	"today";0.038
"open";0.046	"tomorrow";0.036
"air";0.042	"off";0.035

"Regent's Park Open Air Theatre's A Midsummer Night's Dream. Absolutely brilliant.

Topic 94, 0.696	Topic 92, 0.178
"night";0.157	"night";0.080
"dream";0.153	"last";0.070
"midsummer";0.134	"great";0.061
"theatre";0.046	"show";0.040
"open";0.046	"saw";0.040
"air";0.042	"loved";0.036

How can we use these models?

Opinions

"Jamie Parker as Henry V at #TheGlobe was AMAZING. He said 'Cry God for Harry, England & St.Geooorge!' Then we won 6 Gold medals! #Olympics" ok

"Regent's Park Open Air Theatre's A Midsummer Night's Dream. Absolutely brilliant.

Expectations

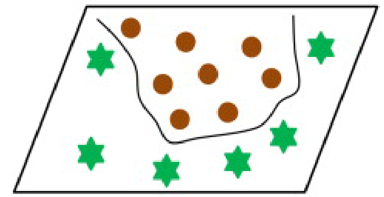
"Tonight I'm going to see a play I've never seen before... 'A Midsummer Night's Dream'."

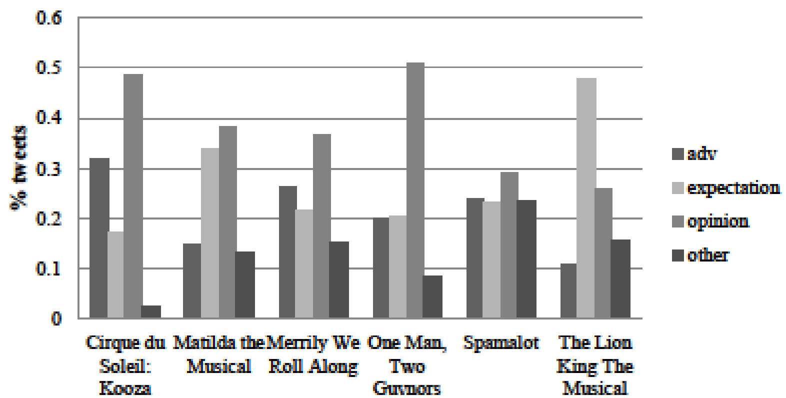
"At the Open Air Theatre for A Midsummer Night's Dream #excited"

Advertisement

"FLASH SALE! Get tickets for 'One Man, Two Guvnors' at The Haymarket Theatre for only £39.99! Buy before time runs out! http://t.co/CSVpgyfP"

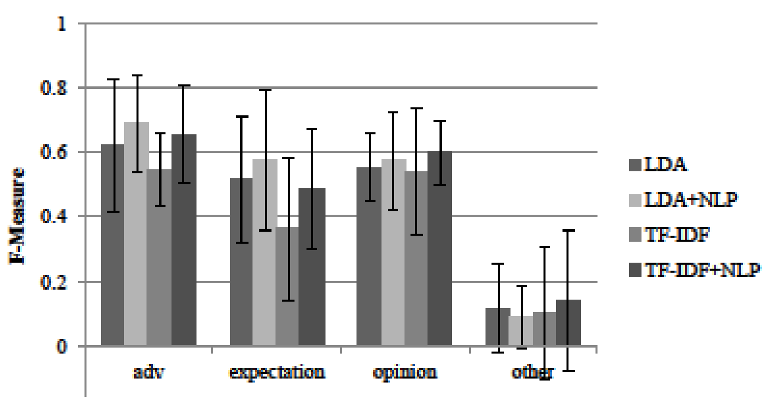
Tweet Classifier





Evaluation of classifiers using precision, recall and f-measure

	P	R	F	P	R	F	P	R	F
		Spamalot			Merrily We Roll Along			Matilda the Musical	
TF-IDF	0.509	0.442	0.444	0.503	0.557	0.493	0.465	0.471	0.435
LDA	0.479	0.496	0.484	**0.634**	**0.608**	**0.611**	0.494	0.563	0.521
TF-IDF + NLP	0.599	0.575	0.573	0.543	0.547	0.524	0.464	0.52	0.49
LDA + NLP	**0.617**	**0.653**	**0.623**	0.565	0.529	0.544	**0.506**	**0.563**	**0.529**
		Cirque du Soleil: Kooza			One Man, Two Guvnors			The Lion King The Musical	
TF-IDF	0.484	0.554	0.488	0.631	0.651	0.597	0.471	0.515	0.423
LDA	**0.627**	**0.635**	**0.608**	0.575	0.583	0.555	0.584	0.621	0.585
TF-IDF + NLP	0.524	0.558	0.5	0.697	0.72	0.695	0.497	0.584	0.526
LDA + NLP	0.594	0.615	0.599	**0.708**	**0.748**	**0.725**	**0.617**	**0.653**	**0.623**



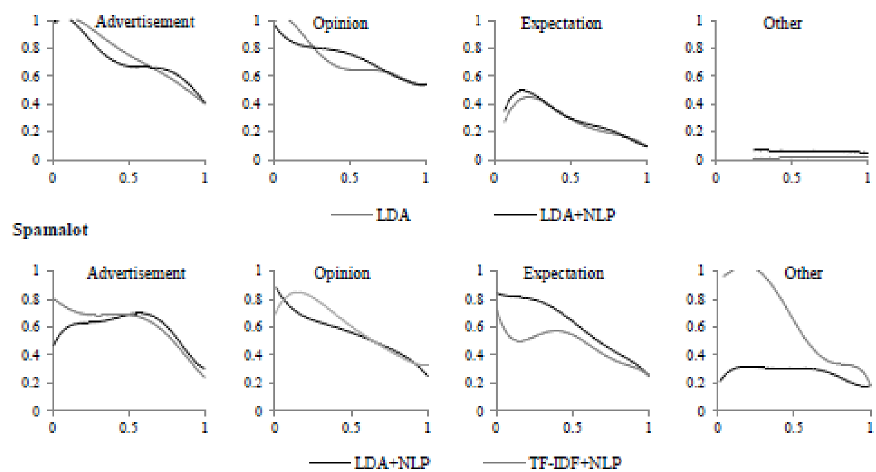


Figure 5. Precision (y-axis) and recall (x-axis) curve per category for two of the shows

Part-Of-Speech Tagger