Latent Theme

Internship
Plan

• Context of the internship
• My mission
• Models used
• My work
• Results
• Problems
• What could have done
• Conclusion
Context

Compare 2 newspapers on the same area

Can we find an answer to the gap of popularity looking at the layout of latent themes?
My first mission

• Scrap articles from the newspaper OUEST-FRANCE into html files
• Convert html files into txt files
• Filter articles creating different models that we will compare (Lemmatization, removing punctuation and common words, lower case...)
• Matrix or vector representation of the documents
• Understanding Python code of pLSA and LDA furnished and adapt them to our data
• Running pLSA and LDA on our different models
• Representing the topics that come out in tables
• Analyze the results
• Repeat all above on the newspaper LETELEGRAMME to compare it with OUEST-FRANCE
• Same for the American newspaper Chicago Tribune and Chicago Sun Time (if time allows it)
• Use latex to write the final report
Models used

Vocabulary = \{one, two, three\}
\[v_d = (1:4, 2:1, 3:0)\]
\[v_d = (1:4, 2:1)\]

**PLSA**

- Each word has an aleatory value
- \(D \in \{1,...,N\}\)
- \(W \in \{1,...,M\}\)
- \(Z \in \{1,...,K\}\)

\[P(d, t) = \sum p(z)p(d, t|z)\]

Parameters estimation (maximization)

Nb tot parameters : \(K+KM+KN\) (overfitting)

**LDA**

Don’t want to be linked to the number of documents

Nb tot parameters : \(K+KM\)
Model used

PLSA

Input: collection $D$, num. of topics $|T|$, num. of iterations $i_{\text{max}}$;
Output: distributions $\phi$, $\theta$;

1. initialize $\phi_{wt}$, $\theta_{td}$ for all $d \in D$, $w \in W$, $t \in T$;
2. for all iterations $i = 1, \ldots, i_{\text{max}}$
3. $n_{wt}, n_{td}, n_t, n_d := 0$ for all $d \in D$, $w \in W$, $t \in T$;
4. for all documents $d \in D$ and terms $w \in d$
5. $p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ for all $t \in T$;
6. $n_{wt}, n_{td}, n_t, n_d := n_{dw}p_{tdw}$ for all $t \in T$;
7. $\phi_{wt} := \frac{n_{wt}}{n_t}$ for all $w \in W$, $t \in T$;
8. $\theta_{td} := \frac{n_{td}}{n_d}$ for all $d \in D$, $t \in T$;

Usually $i_{\text{max}} = 20$. $50$ iterations are sufficient. Time is $O(n|T|i_{\text{max}})$. 

Sara OUNISSI - RICM4 Multi
My work

- Theory
- Scraping (Python - 4243)
- Filter
- Df voc
- Representation (vect – mat)
- PLSA
Results

See htlm files
Problems

- Computer
- LeTelegramme
- Health articles TM-LDA
- LDA
To do

- Run my programs on LETELEGRAMME
- Solve LDA and run for comparison
- Compare with American newspapers
Conclusion

• First experience, interesting subject
• Theory part
• Python
• Autonomous
• Work can be used
Questions?