Topic Modelling

About the Module

- What are Topics

Analytics Vidhya

About the Module

- What are Topics
- Introduction to topic modelling

Analytics Vidhya

About the Module

- What are Topics
- Introduction to topic modelling
- Latent dirichlet allocation

About the Module

- What are Topics
- Introduction to topic modelling
- Latent dirichlet allocation
- Implementation of topic modelling

Analytics Vidhya

Topics

Topics

- A repeating group of statistically significant tokens or words in a corpus

Topics

- A repeating group of statistically significant tokens or words in a corpus
- Statistical Significance

Topics

- A repeating group of statistically significant tokens or words in a corpus

- Statistical Significance
  -Group of words occurring together in the documents

Topics

- A repeating group of statistically significant tokens or words in a corpus

- Statistical Significance
  -Group of words occurring together in the documents
  -Similar term and inverse document frequencies intervals

Topics

- A repeating group of statistically significant tokens or words in a corpus

- Statistical Significance
  -Group of words occurring together in the documents
  -Similar term and inverse document frequencies intervals
  -Frequently occurring together

Topics

- A repeating group of statistically significant tokens or words in a corpus

- Statistical Significance
  -Group of words occurring together in the documents
  -Similar term and inverse document frequencies intervals
  -Frequently occurring together

Topic 1		Topic 2		Topic 3	
term	weight	term	weight	term	weight
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

Analytics Vidhya

Topics

- A repeating group of statistically significant tokens or words in a corpus

- Statistical Significance
  -Group of words occurring together in the documents
  -Similar term and inverse document frequencies intervals
  -Frequently occurring together

Topic 1		Topic 2		Topic 3	
term	weight	term	weight	term	weight
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

Analytics Vidhya

Topics

- A repeating group of statistically significant tokens or words in a corpus

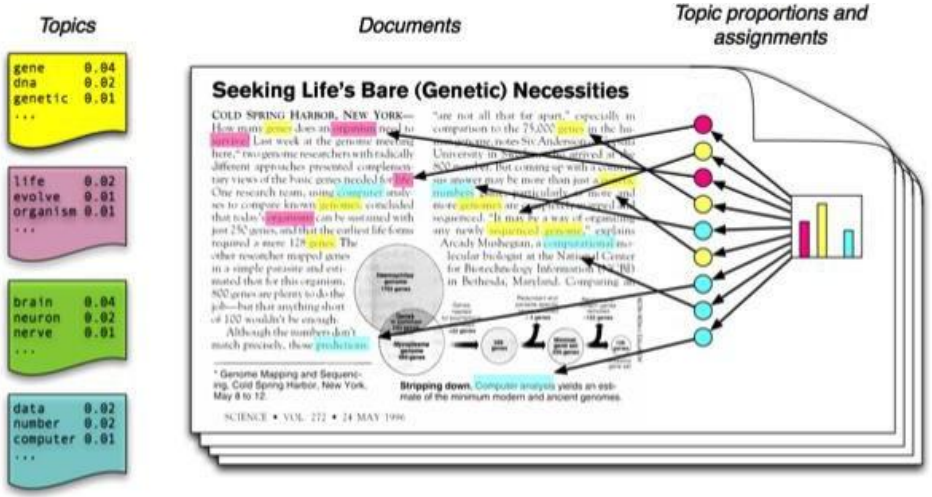
- Statistical Significance
  -Group of words occurring together in the documents
  -Similar term and inverse document frequencies intervals
  -Frequently occurring together

Topic 1		Topic 2		Topic 3	
term	weight	term	weight	term	weight
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

Analytics Vidhya

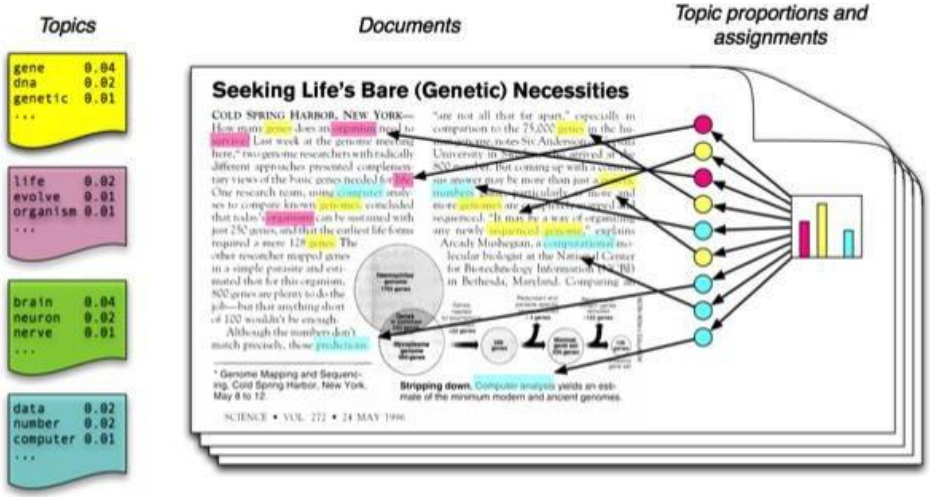
Topic Modelling

Topic Modelling



Analytics Vidhya

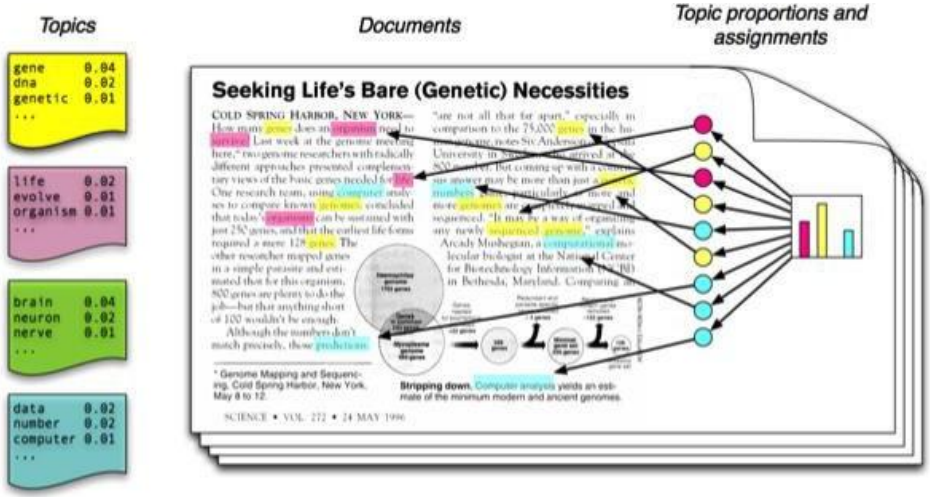
Topic Modelling



- Process to find the topics form documents in an unsupervised manner

Analytics Vidhya

Topic Modelling



- Process to find the topics form documents in an unsupervised manner
- Text mining approach to find recurring patterns in the text documents

Analytics Vidhya

Importance of Topic Modelling

Importance of Topic Modelling

- Document Categorization

Importance of Topic Modelling

- Document Categorization
- Document Summarization

Importance of Topic Modelling

- Document Categorization
- Document Summarization
- Dimensionality Reduction

Importance of Topic Modelling

- Document Categorization
- Document Summarization
- Dimensionality Reduction
- Information Retrieval

Importance of Topic Modelling

- Document Categorization
- Document Summarization
- Dimensionality Reduction
- Information Retrieval
- Recommendation Engines

Topic Modelling Techniques

- LDA – Latent Dirichlet Allocation
- NNMF – Non-Negative Matrix Factorization
- LSA – Latent Semantic Allocation

Latent Dirichlet Allocation

Latent Dirichlet Allocation

Document 1: I want to have fruits for my breakfast.
Document 2: I like to eat almonds, eggs and fruits.
Document 3: I will take fruits and biscuits with me while going to Zoo
Document 4: The zookeeper feeds the lion very carefully
Document 5: One should give good quality biscuits to their dogs

Latent Dirichlet Allocation

Document 1: I want to have fruits for my breakfast.
Document 2: I like to eat almonds, eggs and fruits.
Document 3: I will take fruits and biscuits with me while going to Zoo
Document 4: The zookeeper feeds the lion very carefully
Document 5: One should give good quality biscuits to their dogs

LDA Output

Latent Dirichlet Allocation

Document 1: I want to have fruits for my breakfast.
Document 2: I like to eat almonds, eggs and fruits.
Document 3: I will take fruits and biscuits with me while going to Zoo
Document 4: The zookeeper feeds the lion very carefully
Document 5: One should give good quality biscuits to their dogs

LDA Output

-Topic 1: 30% fruits, 15% eggs, 10% biscuits… (… food)
-Topic 2: 20% lion, 10% dogs, 5% zoo… (… animals)

Latent Dirichlet Allocation

Document 1: I want to have fruits for my breakfast.
Document 2: I like to eat almonds, eggs and fruits.
Document 3: I will take fruits and biscuits with me while going to Zoo
Document 4: The zookeeper feeds the lion very carefully
Document 5: One should give good quality biscuits to their dogs

LDA Output

-Topic 1: 30% fruits, 15% eggs, 10% biscuits… (… food)
-Topic 2: 20% lion, 10% dogs, 5% zoo… (… animals)

-Documents 1 and 2: 100% Topic 1
-Documents 3: 100% Topic 2
-Document 4 and Document 5: 70% Topic 1, 30% Topic 2

Thank You

Analytics Vidhya