Topic Modeling in NLP - Topic modeling - 1

One should look for what is and not what he thinks should be. (Albert Einstein)

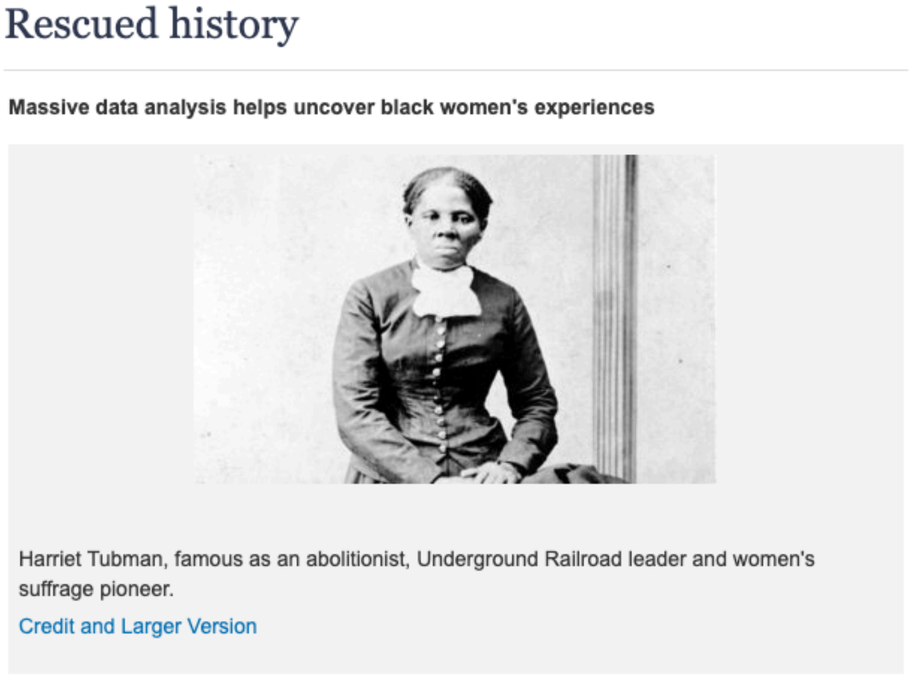
Topic Modeling: Topic introduction

In this part of the course, we will cover the following concepts:

- Topic modeling as an unsupervised method in text analysis
- Latent Dirichlet Allocation as a popular topic modeling algorithm
- Implement LDA on a corpus of documents

Warm up

- More than 80% of the world's data is unstructured
- Applying machine learning to text can reveal insights that go far beyond a count list of words
- Take a few minutes to skim through ***this article***
- What new **insights** were text analysts able to discover that a lifetime of human study couldn't?
- **Share your thoughts** in the chat or aloud



Rescued history

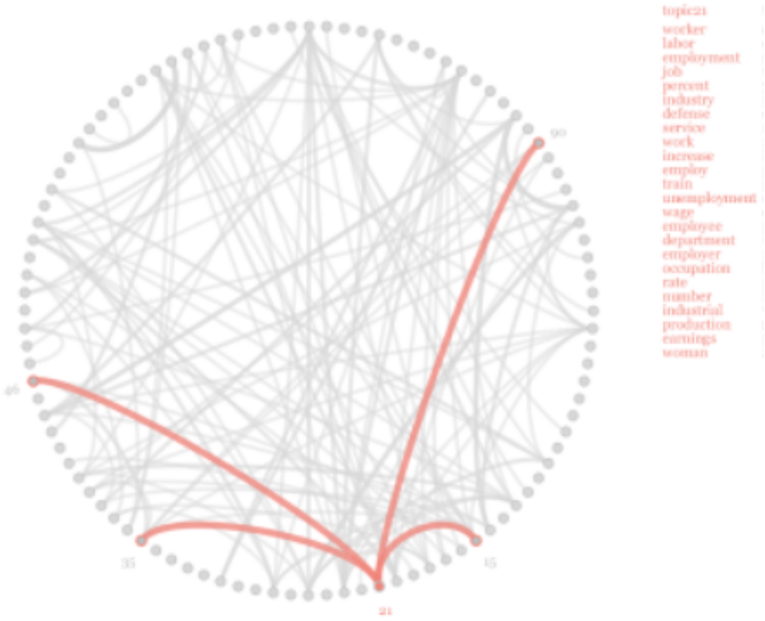
Massive data analysis helps uncover black women's experiences

Harriet Tubman, famous as an abolitionist, Underground Railroad leader and women's suffrage pioneer.

Credit and Larger Version

Rescued history: topic modeling

- Keyword research revealed that many of the documents that referenced topics on the “New Negro” movement also referenced Black suffragettes, revealing that these historical activities were linked
- The finding raises interesting questions about how the two movements, which historians knew were contemporaneous, may have interacted
- This is an example of **topic modeling**, a statistical model that extracts abstract topics from your text based on the frequency of the particular terms used



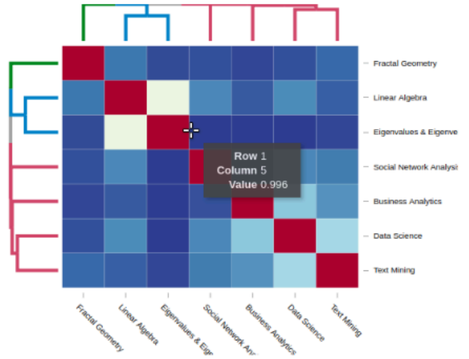
Source

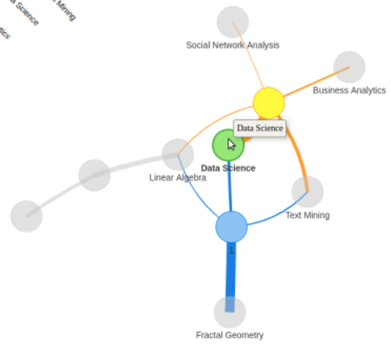
Module completion checklist

Objective	Complete
Summarize the concept of topic modeling	
Describe the process of LDA	

“Bag-of-words” analysis: use cases

- What can be done with a seemingly *crude* approach as the *bag-of-words*?
- Quite a few things, actually! They include:
  - topic modeling
  - word and document similarity query processing
  - word and document clustering
  - sentiment analysis
  - automated document summarization





“Bag-of-words” analysis: `snippet`

- **Topic modeling** is a technique that helps us understand broad topics within our corpus
- We will work with a corpus of `snippet` column from `NYT_article_data.csv`
- We are going to implement topic modeling using a very popular method called **Latent Dirichlet Allocation (LDA)**

LDA: unsupervised text analysis

- How does LDA fall into the category of **unsupervised learning**?
  - **Topics** are formed from **unlabeled** text
  - Clustering documents into “topics” is the basis of this algorithm
  - Clustering is one of the best known unsupervised techniques

`snippet`: topic modeling

- So far, the steps we have taken are:
  - **load** the corpus, where each 'document' is actually one entry in `snippet` column
  - **clean** the text, removing punctuation, numbers, special characters and stop words
  - *stem* the words to their root forms
  - **create** a Document-Term Matrix (DTM) with counts of each word recorded for each document
  - **transform** the DTM to be a weighted term frequency - inverse document frequency matrix

TF-IDF weighted corpus to LDA

- We have our final transformation of our processed documents from `corpus_tfidf`
- The next step is to find out what topics seem to stand out within these documents
  - **Are there groups of documents that all fall under certain topics?**
  - **How can we subset these documents into larger groups?**
- We can find a solution to both of these statements by running an LDA model on the corpus

Module completion checklist

Objective	Complete
Summarize the concept of topic modeling	✔
Describe the process of LDA	

An introduction to LDA

- **Latent Dirichlet Allocation (LDA) is a popular algorithm for topic modeling for many reasons, it allows us to:**
  - reduce dimensionality among large bodies of documents
  - apply other machine learning algorithms to the reduced corpus
  - uncover themes and patterns within your data
- **The algorithm is summarized in three steps:**
  i. Tell the algorithm how many topics you think there are
  ii. Algorithm will assign every word to a temporary topic
  iii. Algorithm will check and update topic assignments
- Here is the ***original paper*** written on the algorithm by David M. Blei, Andrew Y. Ng and Michael I. Jordan

LDA on a simple corpus

- Let’s use a **simple corpus** as an example
- It consists of **three documents** which are actually just three sentences
  - I **ate salad** and **bananas**
  - My brother bought a **cat** and a **hamster**
  - My **cute dog** loves to **munch** on **bananas**

LDA on a simple corpus (cont'd)

- **What do you think LDA will do with these documents?**
- LDA could:
  - Classify the **bold** words under the topic A, which we then might inspect and label **food**
  - Classify the **purple** words under the topic B, which we then might inspect and label **animals**
- LDA is actually defining each of these documents as a bag-of-words and you then label the topics as you see fit

Defining LDA at a document level

- Remember how we applied the TF-IDF transformation to each document?
- This will help you understand why there is a benefit of LDA defining topics on a word level
- **We can infer the content spread of each sentence by a word count:**
  - **Sentence 1:** 100% topic A
  - **Sentence 2:** 100% topic B
  - Sentence 3: 50% topic A, 50% topic B
- **We can derive the proportions that each word constitutes in given topics**
- LDA might produce something like:
  - **Topic A** might comprise words in the following proportions: 40% bananas, 20% ate, 20% salad, 20% munch
  - **Topic B** might comprise words in the following proportions: 25% cat, 25% hamster, 25% dog, 25% cute

LDA in three steps

Let's go back to the three steps of LDA

1. Tell the algorithm how many topics you think there are
2. Algorithm will assign every word to a temporary topic
3. Algorithm will check and update topic assignments

Now, instead of three sentences, let's imagine we have two documents with the following words:

	Document 1		Document 2
	dog		dog
	dog		dog
	cat		hamster
	bananas		munch
	cat		salad

Step 1: number of topics

- **The first step is that we tell the algorithm how many topics you think there are**, this is usually based on:
  - previous analysis
  - informed decision by a subject matter expert
  - random guess
- In trying different estimates, you may pick the one that generates topics to your desired level of interpretability
- In our example, we can probably guess the number of topics by eyeballing the documents, since they are tiny
- **We will guess that there are two topics**

Step 2: algorithm assigns temporary topic

- **The second step is when the algorithm assigns every word in each document to a temporary topic**
  - Topic assignments are temporary, they will be updated in **Step 3**
  - Temporary topics are assigned according to a ***Dirichlet distribution***
  - If a certain word appears twice, it may be assigned to two different topics
- Let's look at how topics have been assigned in our small example, remember we are dealing with topic A and topic B

	Document 1		Document 2
B	dog	?	dog
B	dog	B	dog
B	cat	B	hamster
A	bananas	A	munch
B	cat	A	salad

Step 3: checking and updating topic assignments

- Step 3 is the iterative step of the algorithm, where topics are checked and updated as the algorithm loops through each word in every document
- The algorithm is looking at two main criteria:
  - **a)** How prevalent is the word across topics?
  - **b)** How prevalent are the topics in the document?
- Remember the question marked item in *Document 2* from step 2?
- We will now see how the algorithm iterates and updates the topic for the **?** from step 2, and the assignment for **dog** in document 2

Step 3a: word across topics

How prevalent is the word across topics?

- Dog seems to be prevalent within topic B and not seen in topic
- A dog word picked randomly would more likely be about topic B

	Document 1		Document 2
B	**dog**	**?**	**dog**
B	**dog**	**B**	**dog**
B	cat	B	hamster
A	bananas	A	munch
B	cat	A	salad

Step 3b: topics in the document

How prevalent are the topics in the document?

- Now we look within document 2
- We see that *the words are split 50/50 between* **topic A** and **topic B**
- Therefore, when inspecting the document, **dog** has a 50/50 chance of either topic

	Document 1		Document 2
B	dog	**?**	**dog**
B	dog	**B**	**dog**
B	cat	**B**	**hamster**
A	bananas	**A**	**munch**
B	cat	**A**	**salad**

- We weigh both criteria and see that **dog** from document 2 seems to fit more within **topic B**

What are the topics about?

- **What could topic B be about?**

- **And topic A?**

Step 3c: assign topic

- The process that we used to allocate dog to topic B is repeated on each word in each document, and this cycle occurs multiple times
- This iterative updating is the key feature of LDA that allows us to arrive finally at coherent topics

	Document 1		Document 2
B	dog	**B**	**dog**
B	dog	**B**	**dog**
B	cat	**B**	**hamster**
A	bananas	**A**	**munch**
B	cat	**A**	**salad**

Knowledge check

Module completion checklist

Objective	Complete
Summarize the concept of topic modeling	✔
Describe the process of LDA	✔

Congratulations on completing this module!