**Topic Modeling**

**What is topic modeling?:**

It is a probabilistic model which contains information about topics in the text. For example, if we are working with a corpus of newspaper articles, possible topics would be weather, politics, sport, and so on.

**Why would such topic models be important in the world of text processing?**

Traditionally, information retrieval involved using **words** to identify **similarity** or **relevance** - **now**, we can instead search and arrange our files **more broadly**, with **topics** instead of words.

Topics are distribution of words – a probabilistic distribution of words. Since we know the words and count of the words in documents, we can use this knowledge to generate these topic models. Once we have our topic model, we can start representing all our documents as topic distributions!

For example, the topic that we would call the weather topic in the newspaper corpus would just be a collection of words (such as **sun, temperature, wind, storm, and forecast**), with the associated **probability of those words appearing in the topic**.

In the newspaper corpus, instead of **clustering based on TF-IDF or bag-of-words**, we can now cluster according to the topics. We can also explore the documents in each topic, to better understand the topics, or themes.

Creating topic models for your text corpus is also useful when we want to **explore our dataset**, to see what kind of documents our corpus contains, by just observing the topics.

By arranging our documents in chronological order, we can further see how documents in a topic evolved over time. **Why is this interesting, or useful?**

When time-arranged documents from the research journal **science** were **topic modeled** keeping time-stamps in mind (a technique called Dynamic Topic Modeling), the results were fascinating.

The topic that we associated with **atomic physics** started in **1881** with a high chance of finding the word matter, motion, and light. By the year **1999**, these words under the same topic soon became state, energy, and electron!

**However, a topic is merely a probabilistic distribution of words, and doesn't create its own label, or title.**

**Topic modelling with genism:**

We will be using Gensim [ 2 ] to create our models, which has implementations of **Latent** **Dirichlet Allocation (LDA), Latent semantic analysis (LSA), Hierarchical Dirichlet Process (HDP), and Dynamic Topic Modelling (DTM)** to help us with this.

All of these algorithms have a few things in common – they **assume** words in documents have underlying probabilistic distributions and attempts to find out these distributions. **These distributions end up being our topics.**

The way we attempt to identify these distributions (which is with mathematical and statistical techniques) is what makes these algorithms different.

**Example: Latent Dirichlet allocation**

The first value in the tuple is the topic id.

(5,

u'0.008\*"israeli" + 0.006\*"palestinian" + 0.005\*"force" + 0.004\*"fire" + 0.004\*"people" + 0.004\*"kill" + 0.004\*"government" + 0.004\*"police" + 0.004\*"day" + 0.004\*"australia"')

***What does this mean?*** Topic ID 5 is made up of the words **israeli**, **palestinian** , **force** , **fire** , and so on, and these are the ones with the **highest probability** in the topic. The number that the word is multiplied with (such as 0.008 with Israeli), is the **probability of that word appearing in that topic distribution**.

*We can look at the words with the highest probability to understand the theme of our topic.*

**Homework:**

**Extraction de Thèmes à partir d'un Corpus de Documents**

**Objectif :** Utiliser *topic modeling* pour extraire les thèmes principaux d'un corpus de documents en français ou anglais. Vous utiliserez **spaCy** pour le pré-traitement des données et **Gensim** pour la modélisation thématique. Utilisez les différents modèles LDA, LSA,…

Questions :

1. Quels thèmes sont détectés par chaque modèle et correspondent-ils aux attentes ?
2. Si vous ajoutez plus de documents, voyez-vous une amélioration de la cohérence des thèmes ?
3. Expérimentez avec des valeurs de num\_topics. Comment cela affecte-t-il la cohérence des thèmes ?