# Bonus Project

Automatic document organization, topic extraction, information retrieval and filtering all have one thing in common. They require text clustering (sometimes also known as document clustering) to be done quickly and accurately.

## Text clustering definition

First, let’s define text clustering. Text clustering is the application of cluster analysis to text-based documents. It uses machine learning and natural language processing (NLP) to understand and categorize unstructured, textual data.

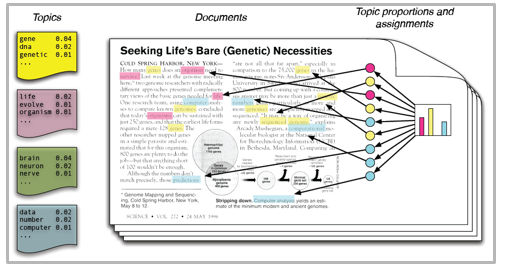
## Hierarchical clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

## Topic Modelling

Topic modeling is an unsupervised machine learning technique that’s capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. Unlike clustering, topic models usually map the documents to topics via a membership function (fuzzy cluster) where one document can consist of multiple topics. For example, *“The food was great, however the parking space was a nuisance”* contains two evident topics (aspects) of “food” and “parking”.

Topic Models are very useful for the purpose of document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection. For Example – New York Times are using topic models to boost their user – article recommendation engines. Various professionals are using topic models for recruitment industries where they aim to extract latent features of job descriptions and map them to the right candidates. They are being used to organize large datasets of emails, customer reviews, and user social media profiles.



# Problem Statement

Given an Excel document about a collection of product reviews on a set of products collected from sources. As a student, you have to devise an unsupervised machine learning algorithm that clusters product reviews into various aspects (for example, price, utility, smell, etc). It is a text mining based problem. It is expected from the students that they design a 2 level hierarchical clustering.

**For example :**

The statement

*“It helps with arthritis pain for a short period of time while making your health deteriorate permanently.”*

Can be classified as into a cluster of “arthritis pain” which again belongs to a cluster of “pain and spasm” related topics.

The same statement can be again clustered based on the “time” being talked about here, hence it further belongs to a sub cluster of “short period”.

Another example can be how the statement:

*“scent isn’t bad value for $ is great, i’ve seen this 2 pack for $45 dollars and when i had a prescription for it it was over $50 for one tube.”*

Which can again be clustered in the category of “smell” with a subcategory of “bad smell”.

Similarly since the price is also discussed in this category, it can again be categorized in the “price” category with its subcategory being the “good value”.

Beware of certain oxymoronic sentences like:

*“It was pretty bad for the price”*

Since the statement has pretty and bad in the same sentence for price there are chances that there will be clusters of “price” with sub categories as “bad value” and “good value”.

There are in total 71,674 such records in the dataset. No submission will get validated and scored, your own review of the results should serve as some notion of cluster validation and topic naming. The final submission will be a piece of code that clusters the full set and names it automatically.