Automated Clustering of Similar Amendments

Jacopo Notarstefano

June 17, 2016

The Problem

The Italian Senate is under a Denial of Service attack.

A senator in the opposition is using software to generate tons of amendments against laws he doesn't want to pass.

This software was originally written for **article spinning**, a black hat SEO technique that generates variations of a text using a regex-like syntax, sometimes called **spintax**.

An Example of Spintax

For example

```
{Hi|Hello}, this is {spin syntax|spintax}.
```

generates

Hi, this is spin syntax.	Hi, this is spintax.	Hello, this is spin syntax.	Hello, this is spintax.

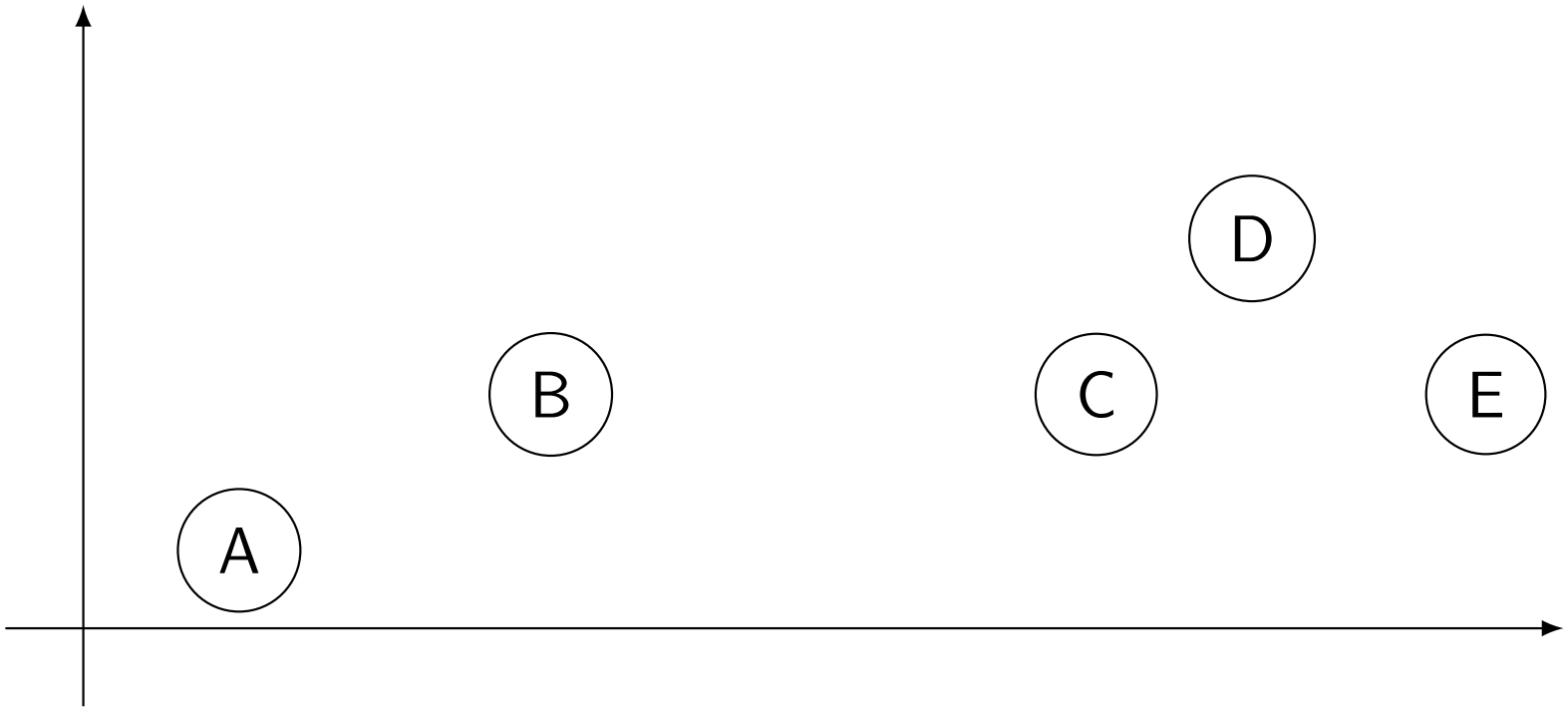
The Extent of the Problem

Solving the Problem

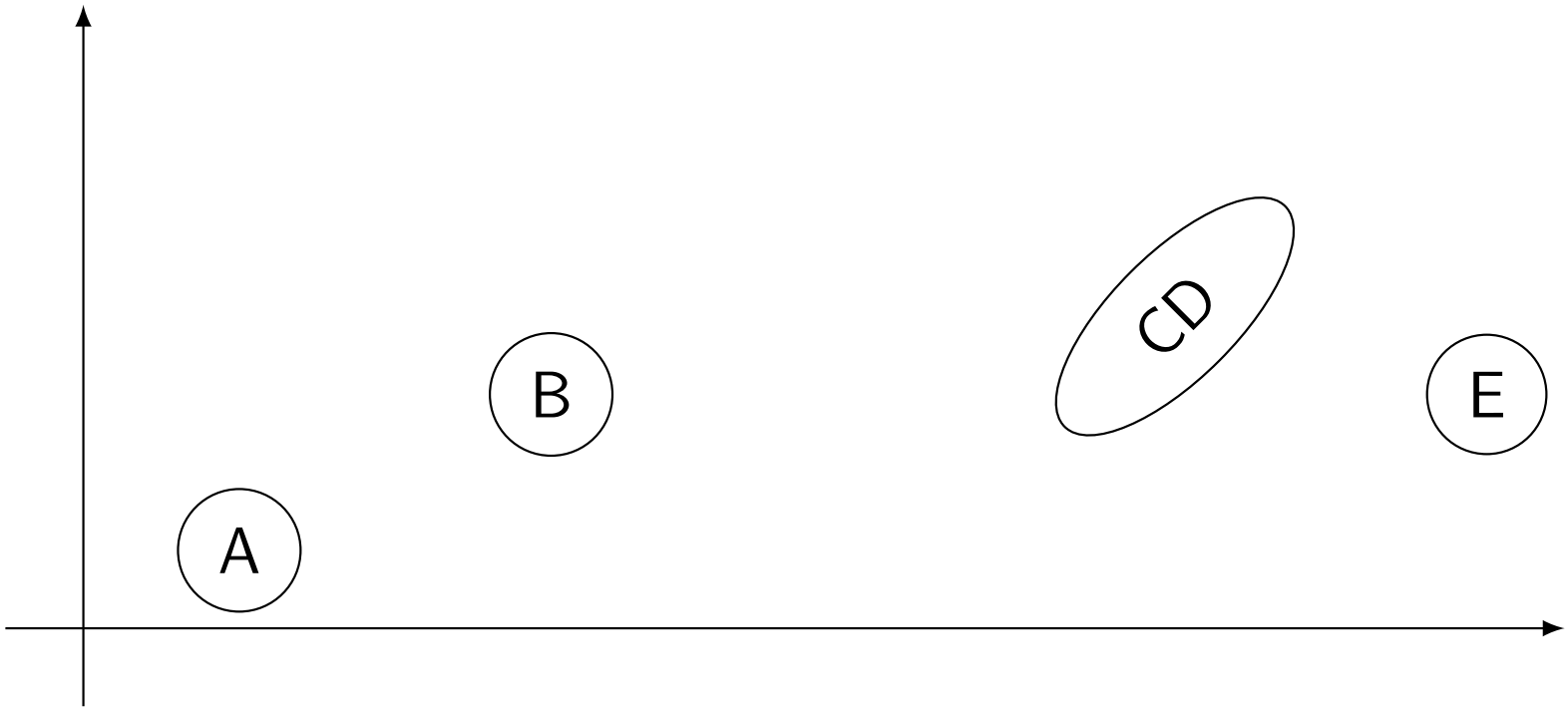
We recognize that the problem we want to solve is an unsupervised clustering in an unknown number of clusters.

The typical algorithm used to solve this problem is called **Hierarchical Agglomerative Clustering**.

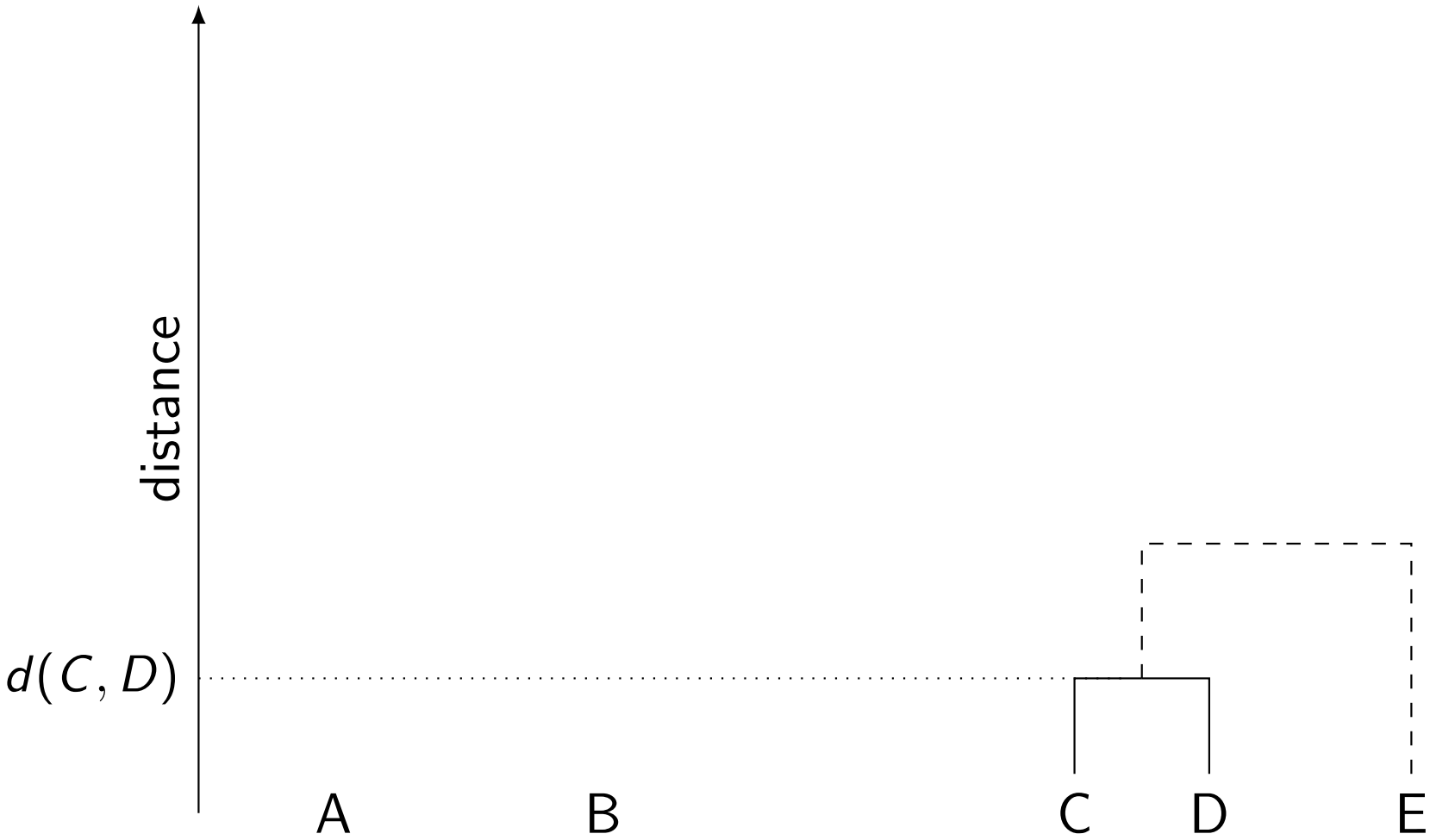
HAC in brief, 1/5



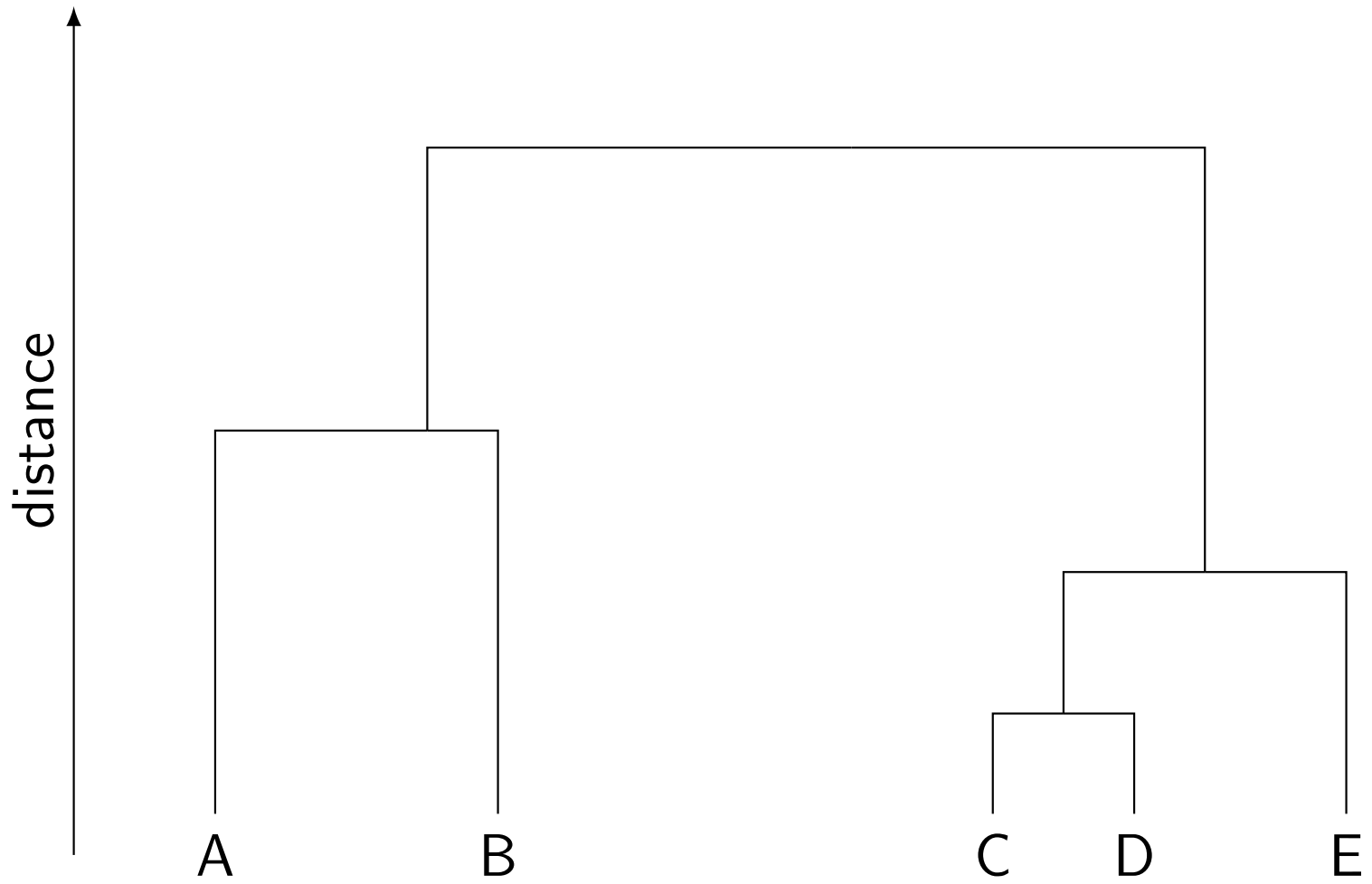
HAC in brief, 2/5



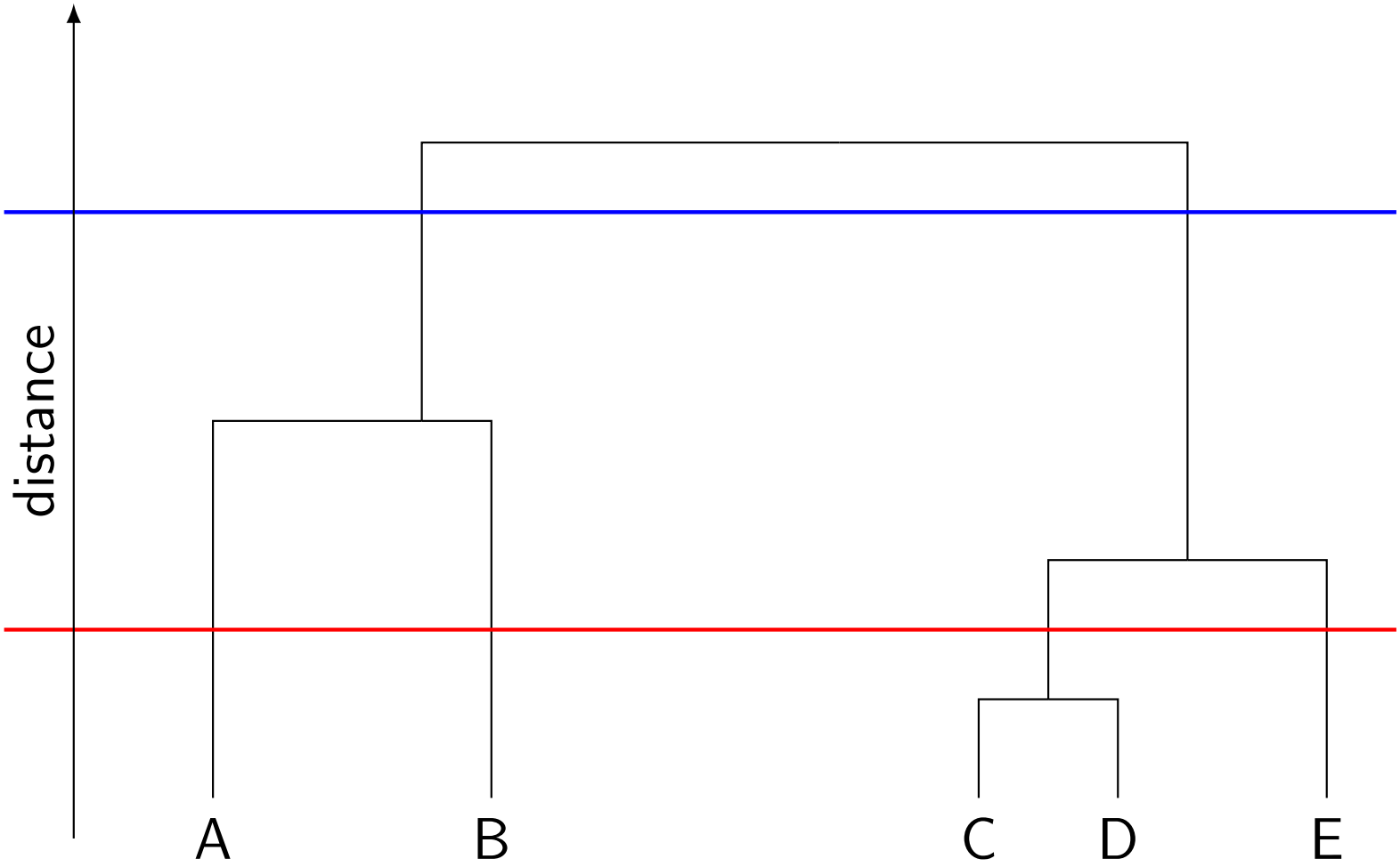
HAC in brief, 3/5



HAC in brief, 4/5



HAC in brief, 5/5



Jaccard's Distance

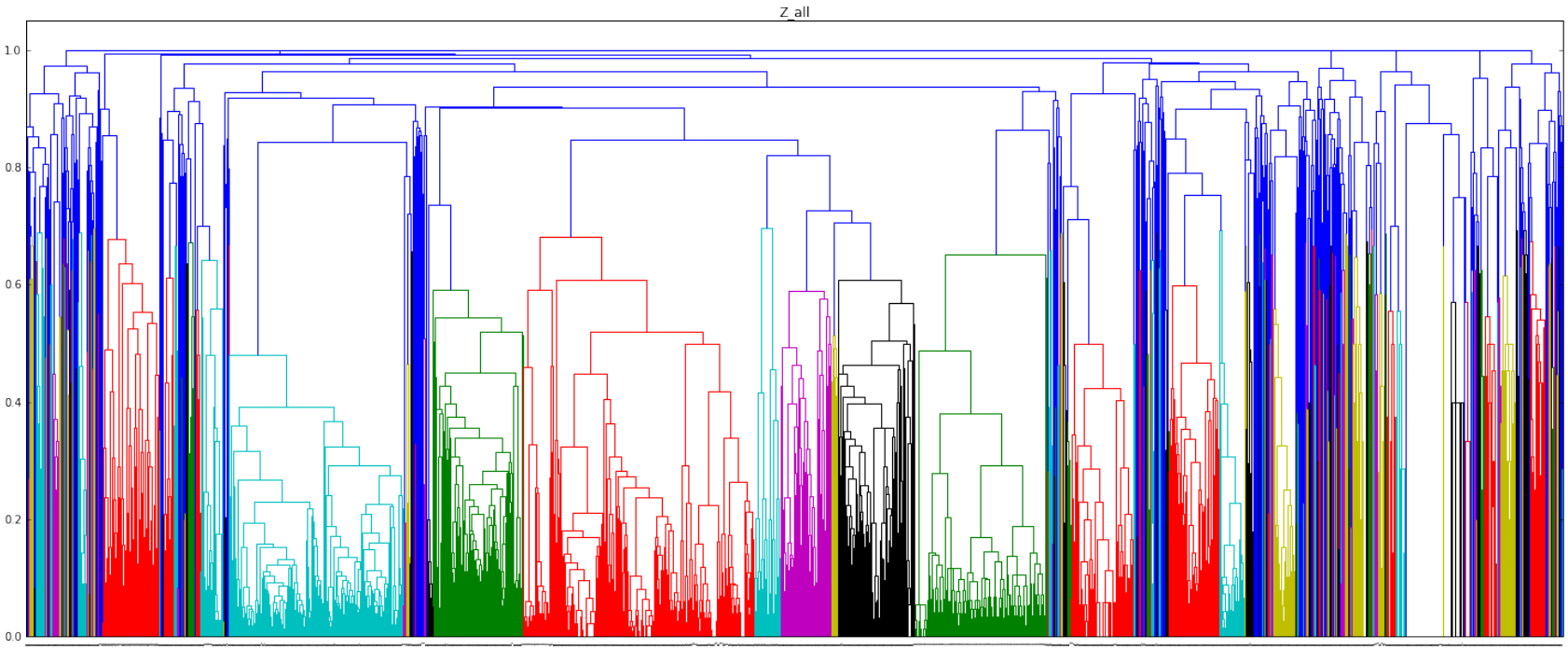
Definition (Jaccard's Distance)

Let A and B be two documents, and let $\texttt{token}(X)$ the function that returns the set of distinct tokens of document X.

Then we call **Jaccard's Distance** the following function:

$$d_{\text{Jaccard}}(A, B) = 1 - \frac{|\texttt{token}(A) \cap \texttt{token}(B)|}{|\texttt{token}(A) \cup \texttt{token}(B)|}$$

The Final Result



More Details

https://github.com/jacquerie/senato.py