

# Document Clustering

## Measuring Document Similarity

| document name               | node+attraction | node# | docs | word#1            | word#2            | word#3            | ... |
|-----------------------------|-----------------|-------|------|-------------------|-------------------|-------------------|-----|
| agricultural_policy_h.....\ |                 | 7     | 1    | consumer          | restoration       | encephalopathy    | ... |
|                             | 53.\            | 333   | 2    | consumer          | labelling         | spongiform        | ... |
| consumer_movement_h...../   |                 | 69    | 1    | consumer          | labelling         | transparency      | ... |
|                             | 43.\            | 379   | 3    | consumer          | spongiform        | encephalopathy    | ... |
| investment_aid_h...../      |                 | 166   | 1    | processing        | encephalopathy    | spongiform        | ... |
|                             | 29-\            | 449   | 4    | consumer          | encephalopathy*   | bovine*           | ... |
| community_control_h...../   |                 | 43    | 1    | monitoring        | ban               | bovine            | ... |
|                             | 22----\         | 471   | 7    | bovine*           | bse*              | consumer*         | ... |
| goat_h.....\                |                 | 143   | 1    | <i>scrapie</i>    | <i>infect</i>     | <i>scientific</i> | ... |
|                             | 62.\            | 304   | 2    | <i>scientific</i> | <i>scrapie</i>    | veterinary        | ... |
| press_h...../               |                 | 196   | 1    | <i>scientific</i> | bovine            | veterinary        | ... |
|                             | 42..../         | 387   | 3    | <i>scrapie</i>    | <i>scientific</i> | <i>infect</i>     | ... |
| cosmetic_product_h...../    |                 | 74    | 1    | <i>scrapie</i>    | encephalopathy    | <i>infect</i>     | ... |

Small sample cluster of seven documents and the first three of a ranked list of indexing words for each document.

The system also calculates the most representative indexing words for each document cluster.

Clustering of multilingual document collections by using language-independent Eurovoc descriptors as input