Overview: An open source document clustering and search tool

Description

Tweet

â??**Overview** *is an open-source tool originally designed to help journalists find stories in large numbers of documents, by automatically sorting them according to topic and providing a fast visualization and reading interface. Itâ??s also used for qualitative research, social media conversation analysis, legal document review, digital humanities, and more. Overview does at least three things really well.*

- *Find what you donâ??t even know to look for.*
- *See broad trends or patterns across many documents.*
- *Make exhaustive manual reading faster, when all else fails.*

Search is a wonderful tool when you know what youâ??re trying to find â?? and Overview includes advanced search features. Itâ??s less useful when you start with a hunch or an anonymous tip. Or there might be many different ways to phrase what youâ??re looking for, or you could be struggling with poor quality material and OCR error. By automatically sorting documents by topic, Overview gives you a fast way to see what you have .

In other cases youâ??re interested in broad patterns. Overviewâ??s topic tree shows the structure of your document set at a glance, and you can tag entire folders at once to label documents according to your own category names. Then you can export those tags to create visualizations.

Rick Davies Comment: This service could be quite useful in various ways, including clustering sets of Most Significant Change (MSC) stories, or micro-narratives form SenseMaker type exercises, or collections of Twitter tweets found via a key word search. For those interested in the details, and preferring transparency to apparent magic, Overview uses the k-means clustering algorithm, which is explained broadly here. One caveat, the processing of documents can take some time, so you may want to pop out for a cup of coffee while waiting. For those into algorithms, here is a healthy critique of careless use of k-means clustering i.e. not paying attention to when its assumptions about the structure of the underlying data are inappropriate

It is the combination of searching using keywords, and the automatic clustering that seems to be the most useful, to meâ?¦so far. Another good feature is the ability to label clusters of interest with one or more tags

I have uploaded 69 blog postings from my Rick on the Road blog. If you want to see how Overview hierarchically clusters these documents let me know, I then will enter your email, which will then let Overview give you access. It seems, so far, that there is no simple way of sharing access (but I am inquiring).

Category

1. Media
2. Software

3. Software

Date
12/09/2025
Date Created
23/01/2015
Author
admin