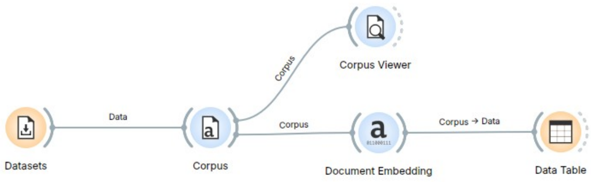
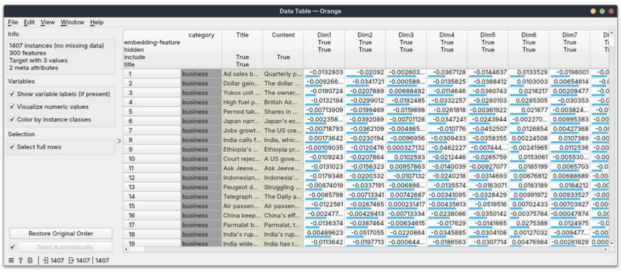
No code lab

Text processing

> *Hint: before doing this exercise you have to install Orange-Text extensions. Open the [Options] menu in the main window, then choose [Add-ons], search for Text and click the checkbox to mark the add-on for installation. Click OK. You will most probably have to restart the tool after add-on installation.*

1. Use **Datasets** operator to load the *BBC3* dataset. This dataset contains over 1400 news articles (titles and contents) classified into several classes based on article topics (sports, politics, entertainment, etc.)
2. Send the data to the **Corpus** operator from the [Text Mining] section. Double-click the operator and make sure that *Used text features* is Content, and *Ignored text features* is Title. Send the output of the **Corpus** operator to the **Corpus Viewer** operator from the [Text Mining] section. Double-click the **Corpus Viewer** to see the articles.
3. First, we will compute the text embedding. Send the output of the **Corpus** operator to the **Document Embedding** operator from the [Text Mining] section. Double-click the **Document Embedding** operator to see the options. Orange Data Mining allows you to use two embeddings, a Sentence BERT (SBERT) and fastText embeddings. Select fastText, make sure that you are using English language embeddings, and use "mean" as the aggregator. FastText computes the embedding for each word, so the embedding of the article will be the mean of all text embeddings. Send the output of the **Document Embedding** operator to the **Data Table** and look at the embeddings. Now you have a very good representation of the text data and you could use this representation to do document clustering and document classification.





Info

1407 instances (no missing data)
300 features
Target with 3 values
2 meta attributes

Variables

✓ Show variable labels (if present)

✓ Visualize numeric values

✓ Color by instance classes

Selection

✓ Select full rows

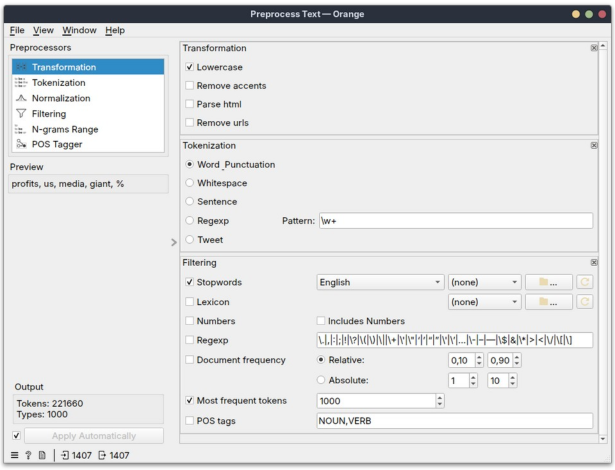
Restore Original Order

✓ Send Automatically

1407 | 1407 | 1407

	category	Title	Content	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Di
embedding-feature				True	True	True	True	True	True	True	T
hidden				True	True	True	True	True	True	True	T
include		True	True								
title		True									
1	business	Ad sales b...	Quarterly p...	-0.0132803	-0.02092	-0.002603...	-0.0367128	-0.0144637	0.0133529	-0.0198001	-0.0
2	business	Dollar gain...	The dollar ...	-0.009266...	-0.0341721	-0.000589...	-0.0135825	-0.0388412	0.0103003	0.00654614	-0.0
3	business	Yukos unit ...	The owner...	-0.0190724	-0.0207889	0.00688492	-0.0114646	-0.0360743	0.0218217	0.00209477	-0.0
4	business	High fuel p...	British Air...	-0.0132194	-0.0299012	-0.0192485	-0.0332257	-0.0280103	0.0285305	-0.030353	-0.0
5	business	Pernod tak...	Shares in ...	-0.00713909	-0.0199469	-0.0119898	-0.0261818	-0.00361922	0.021877	-0.003824...	-0.0
6	business	Japan narr...	Japan's ec...	-0.002358...	-0.0392089	-0.00701128	-0.0347241	-0.0243944	-0.002270...	0.00995383	-0.00
7	business	Jobs growt...	The US cre...	-0.00718793	-0.0362109	-0.004865...	-0.010776	-0.0452507	0.0126854	0.00427368	-0.0
8	business	India calls f...	India, whic...	0.00173642	-0.0230194	-0.0096956	-0.0309433	-0.0358355	0.00224508	0.0107389	-0.00
9	business	Ethiopia's ...	Ethiopia pr...	-0.00109035	-0.0120476	0.000327132	-0.0462227	-0.007444...	-0.00241965	0.0112536	-0.00
10	business	Court rejec...	A US gove...	-0.0109243	-0.0207864	0.0102593	-0.0212446	-0.0265759	0.0153061	-0.005530...	-0.00
11	business	Ask Jeeve...	Ask Jeeve...	-0.0131023	-0.0156323	0.00957863	-0.0140039	-0.00927077	0.0185199	0.0065703	-0.0
12	business	Indonesian...	Indonesia'...	-0.0179348	-0.0200332	-0.0107132	-0.0240218	-0.0314693	0.00676812	0.00688689	-0.00
13	business	Peugeot d...	Struggling ...	-0.00874019	-0.0337191	-0.006898...	-0.0135574	-0.0163071	0.0183189	0.0184212	-0.00
14	business	Telegraph ...	The Daily a...	-0.0085798	-0.00713341	0.00742687	-0.00341095	-0.0328429	0.00991972	0.00933527	-0.00
15	business	Air passen...	Air passen...	-0.0122561	-0.0267465	0.000231417	-0.00435613	-0.0519516	0.00702433	-0.00703927	-0.00
16	business	China keep...	China's eff...	-0.002477...	-0.00429413	-0.00713334	-0.0238086	-0.0350142	-0.00375784	-0.00047874	0.00
17	business	Parmalat b...	Parmalat, t...	-0.0136374	-0.0387464	-0.00634615	-0.017629	-0.0141665	0.0275388	0.0124975	-0.0
18	business	India's rup...	India's rup...	0.00489623	-0.0517055	-0.0220864	-0.0345885	-0.0304108	0.00127032	-0.009477...	-0.00
19	business	India wide...	India has r...	-0.0113642	-0.0197713	-0.000644...	-0.0186563	-0.0307714	0.00476984	-0.00261829	0.000

4. We will return to a much simpler text representation which is a bag-of-words model. Send the output of the **Corpus** operator to the **Preprocess Text** operator from the [Text Mining] section. Double-click on this operator, as you can see, a lot of operations can be performed here. Make sure that you tranform all words to lowercase. Tokenization of the text can be performed by Word_Punctuation. Use stop-word filtering (the operator will remove very popular and common words such as "the", "in", "for", etc.) To speed up computations check the *Most frequent tokens* checkbox and leave only 1000 most frequent words. Send the output of the **Preprocess Text** operator to the **Word Cloud** operator from the [Text Mining] and see the results.



Preprocess Text — Orange

File View Window Help

Preprocessors
- Transformation
- Tokenization
- Normalization
- Filtering
- N-grams Range
- POS Tagger

Preview: profits, us, media, giant, %

Transformation
- ☑ Lowercase
- □ Remove accents
- □ Parse html
- □ Remove urls

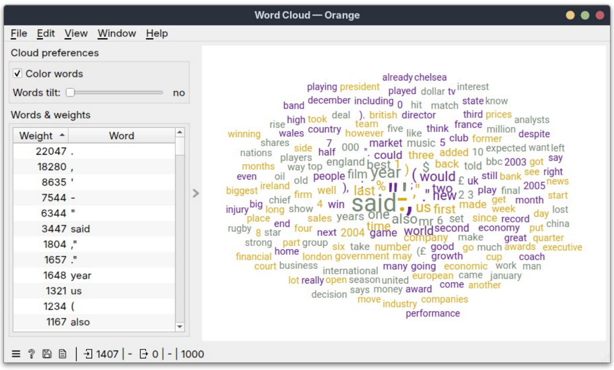
Tokenization
- ● Word_Punctuation
- ○ Whitespace
- ○ Sentence
- ○ Regexp — Pattern: \w+
- ○ Tweet

Filtering
- ☑ Stopwords — English — (none)
- □ Lexicon — (none)
- □ Numbers — □ Includes Numbers
- □ Regexp — \.|,|:|;|!|\?|\(|\)|\||\+|\'|\"|'|'|"|"|\'|\'|...|\-|–|—|\$|&|*|>|<|\/|\[|\]
- □ Document frequency — ● Relative: 0,10 0,90 — ○ Absolute: 1 10
- ☑ Most frequent tokens — 1000
- □ POS tags — NOUN,VERB

Output
Tokens: 221660
Types: 1000

☑ Apply Automatically

1407 1407



Word Cloud — Orange

File Edit View Window Help

Cloud preferences
- ☑ Color words
- Words tilt: no

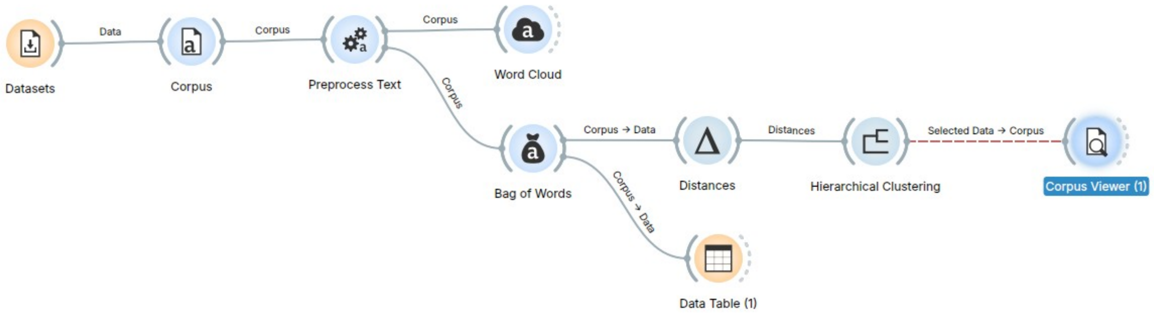
Words & weights

Weight	Word
22047	.
18280	,
8635	'
7544	-
6344	"
3447	said
1804	,"
1657	."
1648	year
1321	us
1234	(
1167	also

1407 | - 0 | - | 1000

already chelsea
playing president played dollar tv interest
band december including 0 hit match state know
rise high took deal). british director third prices analysts
winning wales country team however five like think france million despite
shares side 7 000 market music 5 club former
nations players half could three added 10 expected want left
months way top england best 1 $ back told bbc 2003 got say
even oil old people film year (would £ uk still bank see right
biggest ireland firm well). last % two new 2 3 play final 2005 news
injury big chief long show 4 win said us first made get month start
place end sales years one also mr 6 set since record week day lost
rugby 8 star four next 2004 time game world second economy put china
strong part group six take number company make great quarter
financial home london government may (£ good growth go much awards executive cup coach
court business international many going economic work man
lot really open season united european came january
decision says money award come another
move industry companies
performance

5. Send the output of the **Corpus** operator to the **Bag of Words** operator from the [Text Mining] section. This operator transforms each article to a vector, where each position represents a word and the number represents the number of occurrences of that word in the article. If you want, you can use the **Data Table** operator to look at the **Bag of Words** transformation.
6. Send the **Bag of Words** output to the **Distances** operator from the [Unsupervised] section. This operator will compute the similarity between articles based on the number of co-occurrences of words.
7. Next, send the computed distances to the **Hierarchical Clustering** operator from the [Unsupervised] section. Finally, send the output of the **Hierarchical Clustering** operator to a **Corpus Viewer** operator. This operator will always display documents that are in the currently selected cluster. Your final workflow should be as follows:



8. Double-click the **Hierarchical Clustering** operator and select one of the clusters. Examine the contents of the cluster in the **Corpus Viewer**. Do all articles from the selected cluster share the same category?

Assignment

Repeat steps 6-8 for the document representation computed from word embeddings. Compute the distances between embeddings, run hierarchical clustering and examine the clusters of documents.