The North Face ecommerce

Machine Learning

Context

- **Problem**: E-commerce platform wants to boost its online sales
- **Solution 1**: deploy a recommender system for the products catalogue
- **Solution 2**: Improve the catalogue structure using topic extraction
- **Dataset**: 500 product descriptions gathered from the website

Context

- **Problem**: E-commerce platform wants to boost its online sales
- **Solution 1**: deploy a recommender system for the products catalogue
- **Solution 2**: Improve the catalogue structure using topic extraction
- **Dataset**: 500 product descriptions gathered from the website

Tasks

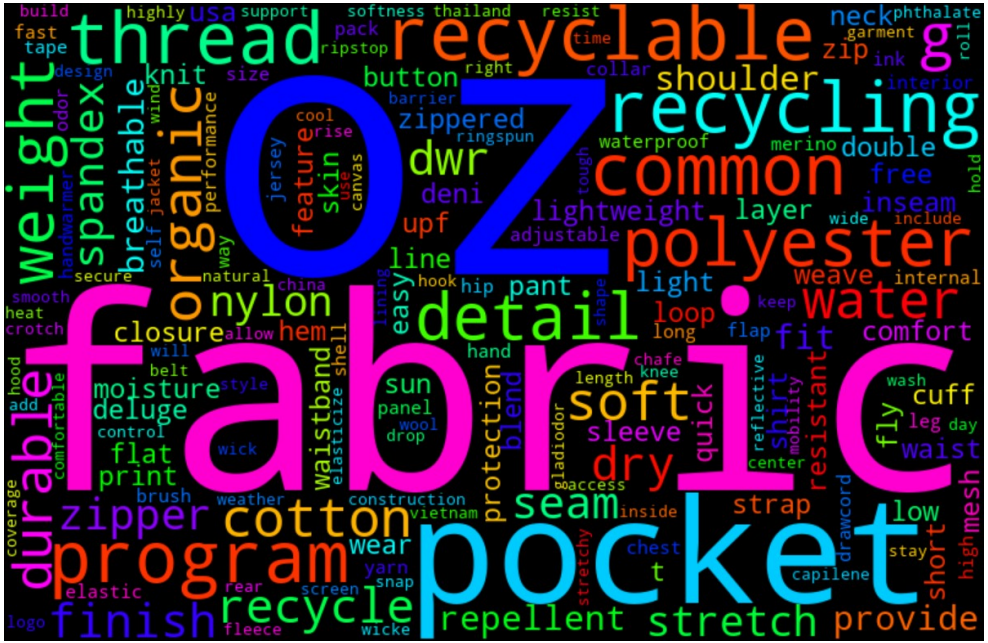
- Process the product descriptions into a TF-IDF matrix
- Use unsupervised learning to make clusters of similar products
- Design a recommender system to find items similar to the selection
- Extract topics from descriptions using latent semantic analysis

Descriptions overview

- 500 product descriptions
- About 100 words each
- Gathered by web scraping
- Corpus preprocessing
  - Remove HTML tags
  - Remove punctuation
  - Remove numbers
  - Remove stop words
  - Strip diacritics and casefold
  - Lemmatize words

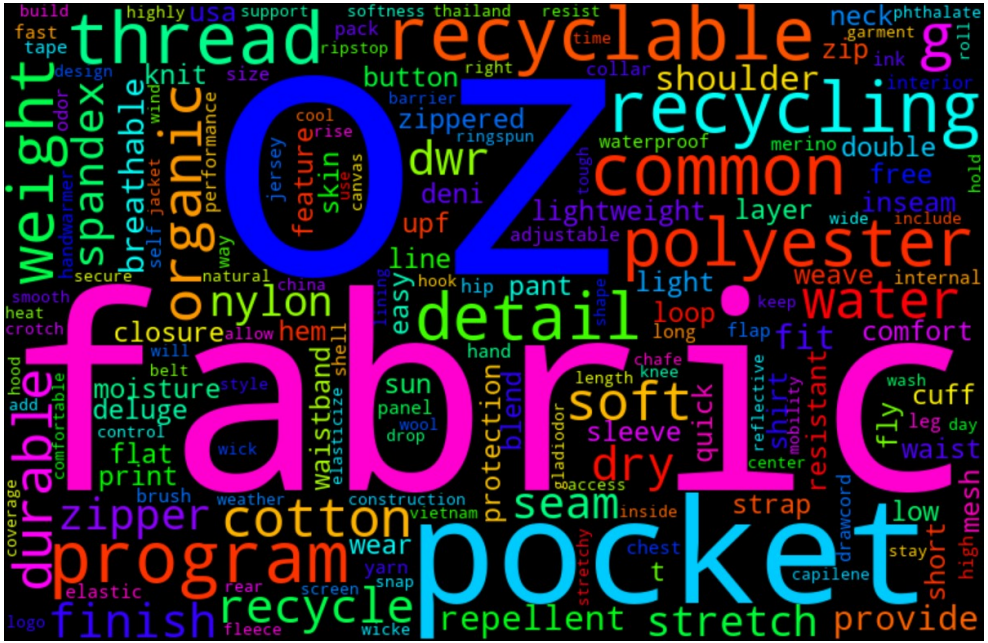
Descriptions overview

- 500 product descriptions
- About 100 words each
- Gathered by web scraping
- Corpus preprocessing
  - Remove HTML tags
  - Remove punctuation
  - Remove numbers
  - Remove stop words
  - Strip diacritics and casefold
  - Lemmatize words



Descriptions overview

- 500 product descriptions
- About 100 words each
- Gathered by web scraping
- Corpus preprocessing
  - Remove HTML tags
  - Remove punctuation
  - Remove numbers
  - Remove stop words
  - Strip diacritics and casefold
  - Lemmatize words



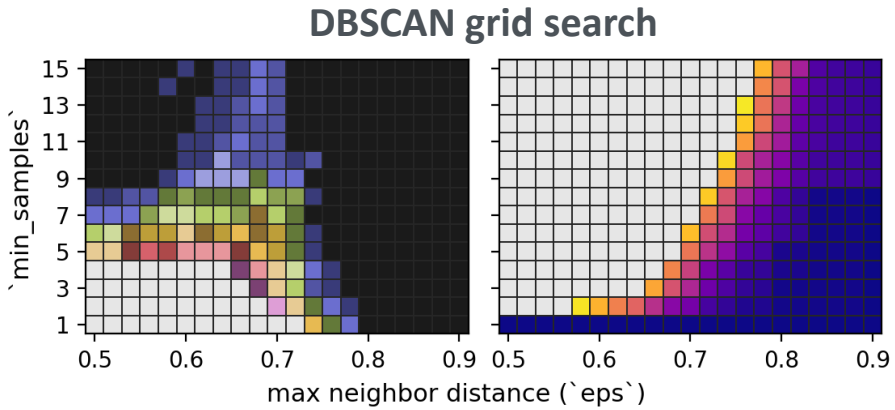
Garment descriptions

Clustering : Methods

- Convert description corpus to TF-IDF
- 2 methods: DBSCAN / HDBSCAN
- Optimize parameters
  - Limit the number of outliers
  - Make about 10 clusters

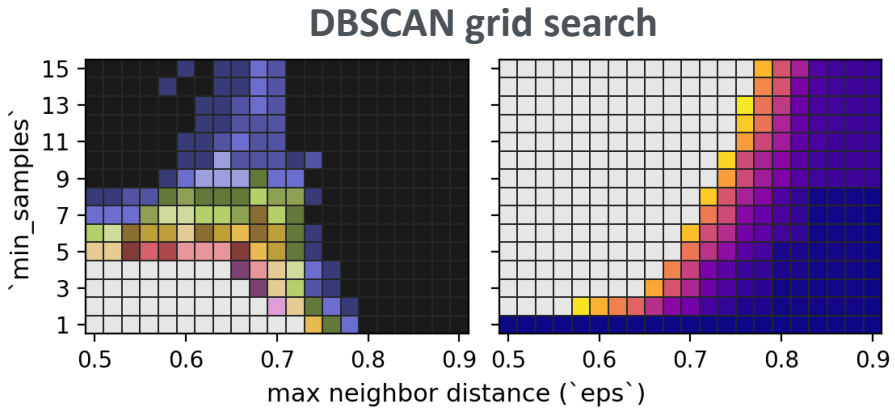
Clustering : Methods

- Convert description corpus to TF-IDF
- 2 methods: DBSCAN / HDBSCAN
- Optimize parameters



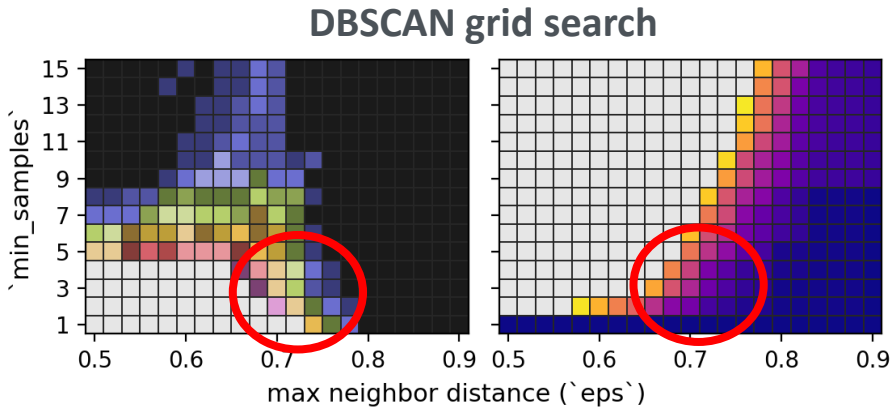
Clustering : Methods

- Convert description corpus to TF-IDF
- 2 methods: DBSCAN / HDBSCAN
- Optimize parameters
  - Limit the number of outliers
  - Make about 10 clusters



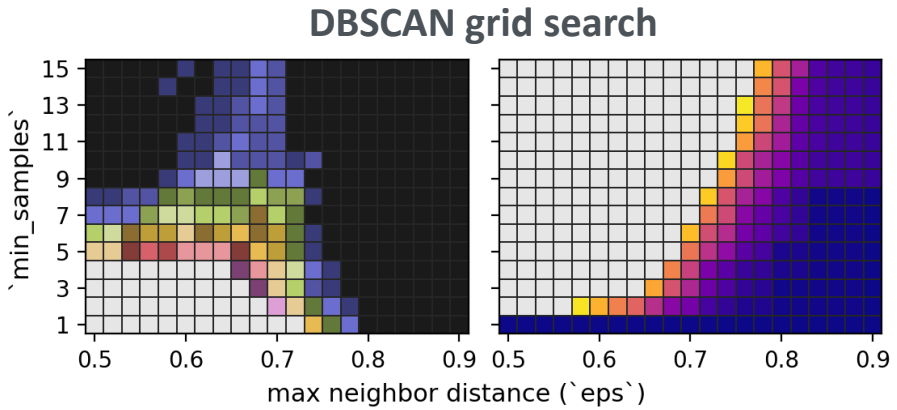
Clustering : Methods

- Convert description corpus to TF-IDF
- 2 methods: DBSCAN / HDBSCAN
- Optimize parameters
  - Limit the number of outliers
  - Make about 10 clusters



Clustering : Methods

- Convert description corpus to TF-IDF
- 2 methods: DBSCAN / HDBSCAN
- Optimize parameters
  - Limit the number of outliers
  - Make about 10 clusters

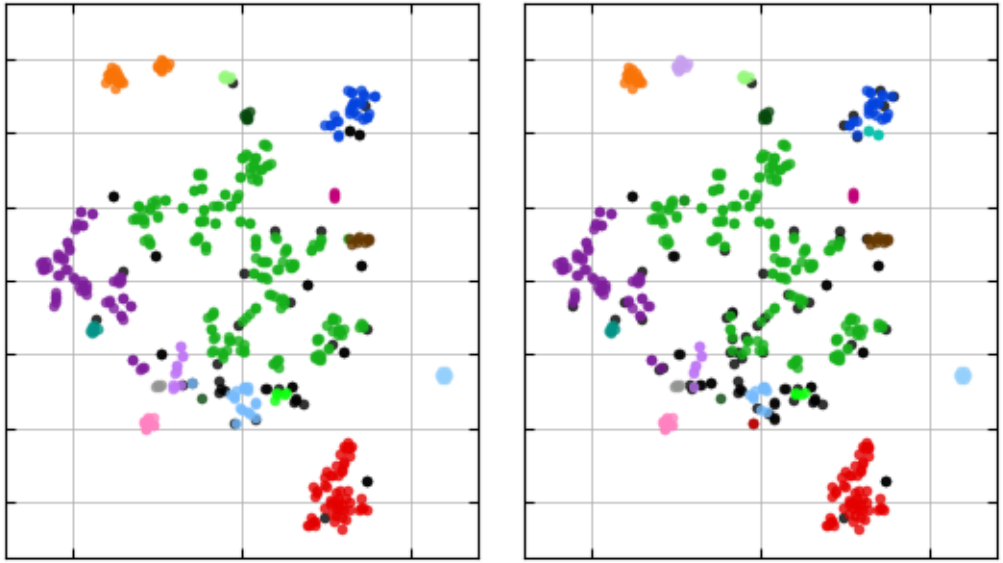


- Difficult to find good parameters!
- DBSCAN
  - 16 clusters
  - 48 outliers (10%)
- HDBSCAN
  - 18 clusters
  - 80 outliers (16 %)
- Strong cluster inhomogeneity
  - Largest cluster : about 180 items
  - Smallest cluster : 4 items

Clustering : t-SNE Vizualization

Low dimensional embedding

t-distributed Stochastic Neighbor Embedding (t-SNE)

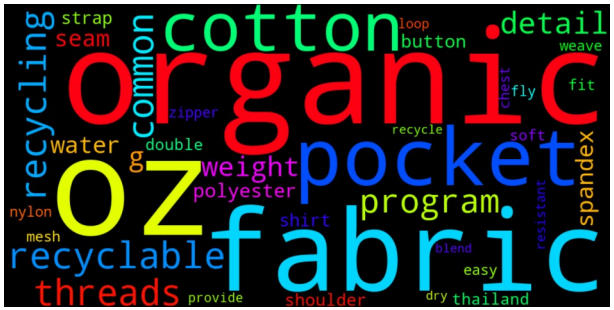


- Similar clusterization with DBSCAN and HDBSCAN
- Small well-defined clusters
- Large central aggregate
  - One big cluster (180 items)
  - Intertwined with outliers

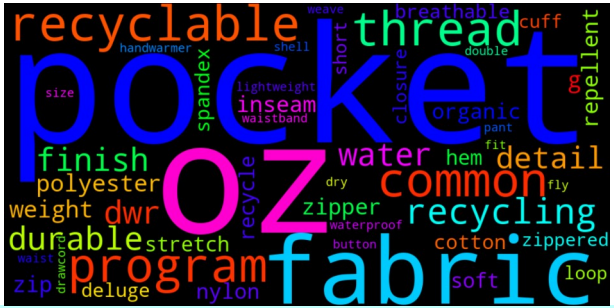
Clustering : Wordcloud visualization

- Clusters difficult to interpret

Outliers



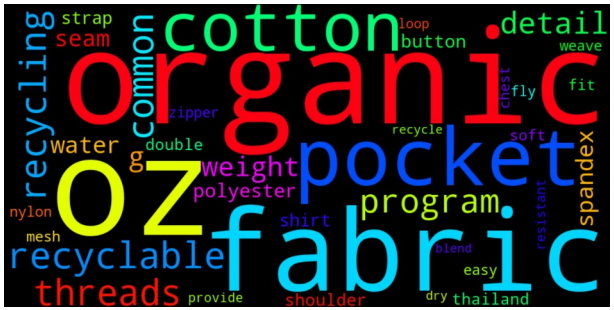
Largest cluster



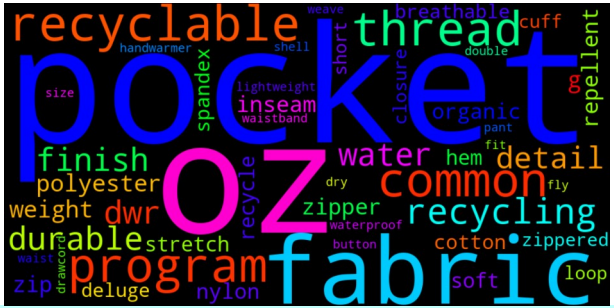
Clustering : Wordcloud visualization

- Clusters difficult to interpret

Outliers

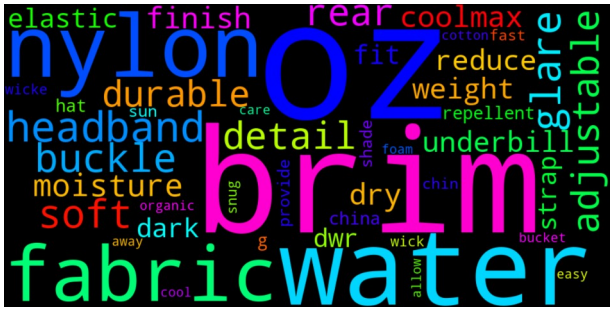


Largest cluster

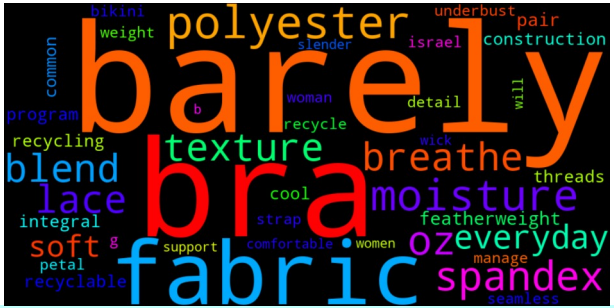


- Clear garment categories

Brims



Bras



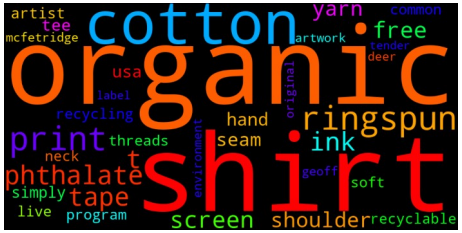
Topic extraction

- Truncated SVD
  - Get eigenvectors of the covariance matrix
- Representation of the corpus topics
- Visualize as wordclouds
  - Split positive and negative components

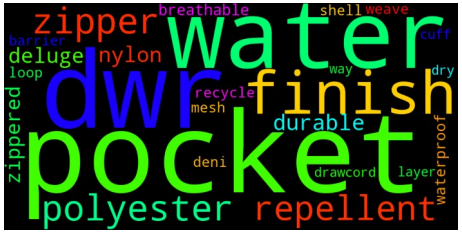
Topic extraction

- Truncated SVD
  - Get eigenvectors of the covariance matrix
- Representation of the corpus topics
- Visualize as wordclouds
  - Split positive and negative components

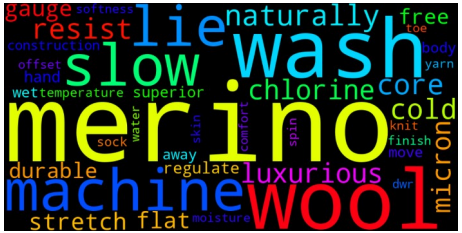
Natural



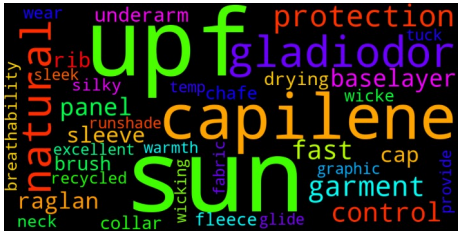
Synthetic



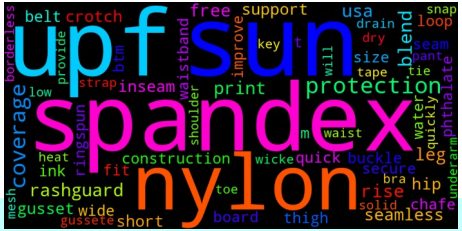
Cold



Hot



Hot / Synthetic



Cold / Organic

Conclusion

- Good clusterization and topic extraction from only 500 descriptions
- Limited clustering performance
- Mixing of topics from different semantic fields
  (eg garment type and garment composition)
- For a better clustering / topic extraction
  - More descriptions
  - Split descriptions (garment usage / composition / ...)

Thanks!