Latent_semarical_analysis

January 19, 2019

```
In [3]: #LSA-Latent semarical analysis
        from sklearn.datasets import fetch_20newsgroups
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.decomposition import TruncatedSVD
        import numpy as np
        categories = ['sci.med', 'sci.space']
        twenty_sci_news = fetch_20newsgroups(categories=categories)
        tf_vect = TfidfVectorizer()
        word_freq = tf_vect.fit_transform(twenty_sci_news.data)
        tsvd_2c = TruncatedSVD(n_components=50)
        tsvd_2c.fit(word_freq)
        np.array(tf_vect.get_feature_names())

        [tsvd_2c.components_[20].argsort()[-10:][::-1]] # it should display words (version iss
```

```
Out[3]: [array([25610, 13553, 13699, 21903, 10113,  6815, 11053,  6668, 24389,
                16985], dtype=int32)]
```

```
In [ ]:
```