Topic Models

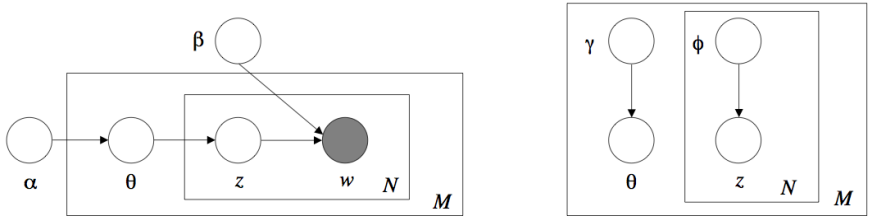
Tecnologías de búsqueda en la web

Marcelo Mendoza

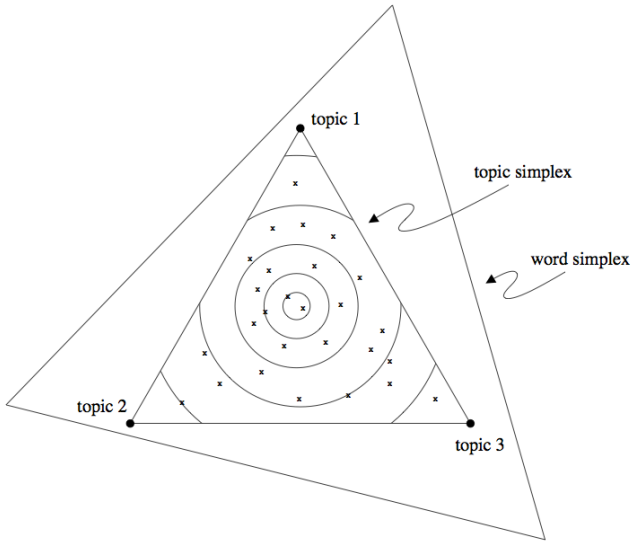
Departamento de Informática
Universidad Técnica Federico Santa María

Topic Models: LDA

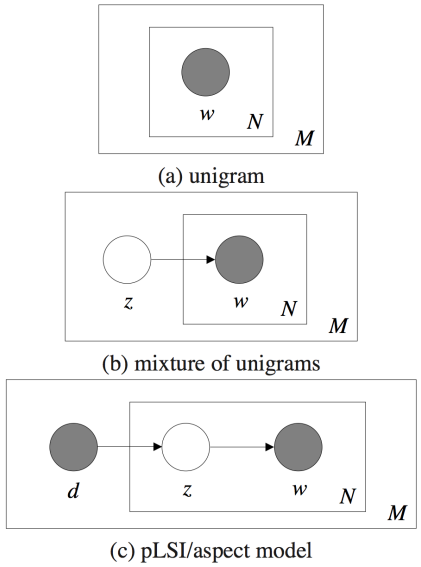
Decoupling trick



LDA: Interpretación geométrica



Modelos Gráficos



Gensim (http://radimrehurek.com/gensim/)

Installing:

```
> sudo pip install --upgrade gensim
```

Corpus (preprocesamiento):

```
> python
>>> from gensim import corpora, models
>>> documents = ["Human machine interface for lab abc computer applications",
...              "A survey of user opinion of computer system response time",
...              "The EPS user interface management system",
...              "System and human system engineering testing of EPS",
...              "Relation of user perceived response time to error measurement",
...              "The generation of random binary unordered trees",
...              "The intersection graph of paths in trees",
...              "Graph minors IV Widths of trees and well quasi ordering",
...              "Graph minors A survey"]
>>> import nltk
>>> from nltk.corpus import stopwords
>>> stop = stopwords.words('english')
>>> texts = [[word for word in document.lower().split() if word not in stop]
    for document in documents]
```


Gensim

Corpus (representación BOW):

```
>>> dictionary = corpora.Dictionary(texts)
>>> corpus = [dictionary.doc2bow(text) for text in texts]
>>> corpora.MmCorpus.serialize('/tmp/deerwester.mm', corpus)
```

LDA fitting:

```
>>> lda = models.LdaModel(corpus, id2word=dictionary, num_topics=2)
```

Term dist. (print_topics(num_topics)):

```
>>> lda.print_topics(2)
```

Top-terms per topic (print_topic(topicid,topn)):

```
>>> lda.print_topic(0, topn=5)
>>> lda.print_topic(1, topn=5)
```

Document dist.:

```
>>> doc2bow = dictionary.doc2bow(texts[1])
>>> lda[doc2bow]
```


LDA update (sobre un nuevo corpus):

```
>> lda.update(new_corpus)
```