Implementation of LDA and Jargon Distance algorithms:

In order to implement the Latent Dirichlet allocation on both the patent data sets and the Toy Data set, ipython was used as the language of choice. The natural language toolkit library and its english language corpus (nltk, nltk corpus (english)) were used in order to come up with stop words, stems and lemmatization.

The general steps followed included the following in order mentioned below.

Tokenization, normalization, removal of stop words, stemming via the porter stemming function and lemmatization.

At this point genism package tools were used to create the new dictionary, corpus and document term frequency matrices. As a general choice, 5 topics were chosen in order to avoid repetition of topics, while maintain a significant distribution of topics. General functions in Gensim such as corpora.Dictionary, doc2bow, Tfidfmodel, ldamodel, and id2word were significantly handy and made LDA easy to use and implement.

The results from LDA implementation on the Patent data set were as mentioned below

```
Top ten words and their frequencies
Topic 1 :
0.063*protector + 0.050*strap + 0.036*section + 0.035*first + 0.032*nozzl
+ 0.032*rotari + 0.028*second + 0.027*jet + 0.026*assembl + 0.020*whirlpoo
l

Topic 2 :
0.078*cm/ + 0.068*accord + 0.068*measur + 0.057*rate + 0.052*method + 0.04
6*flow + 0.042*sleep + 0.042*airflow + 0.039*air + 0.037*area

Topic 3 :
0.109*bag + 0.085*sleep + 0.050*part + 0.050*bodi + 0.039*leg + 0.035*main
+ 0.035*foot + 0.032*substanti + 0.024*user + 0.022*drawstr

Topic 4 :
0.100*edg + 0.058*perineum + 0.050*protect + 0.050*devic + 0.049*gener + 0
.045*panel + 0.042*inch + 0.039*first + 0.035*second + 0.034*loop

Topic 5 :
0.056*member + 0.044*faucet + 0.036*tighten + 0.036*support + 0.031*bathtu
b + 0.026*seat + 0.025*plural + 0.024*defin + 0.023*base + 0.021*connector
```

Observations:

The topics, as such seem to provide a very bleak just by looking at 10 most frequent terms in the topic but important insight about kinds of terms that are present in these papers can be inferred from these majority terms. For example Topic 2 seems to reflect that most of the documents processed are somehow related to measurements and physics. Topic 3 seems to

describing that one topic is related to parts of a person, animal or objects like a table. Topic 5 seems to be related to washroom and toiletries etc. thus LDA seems to be a great tool to infer the hidden structures and similarities between topics covered in variety of documents on the basis of most frequent related words they contain. Since we have no idea at the beginning about the relationships between the documents it might be a good tool to gain basic inference.

The implementation of jargon paper was pretty similar at the beginning to the LDA process and involved use of NLTK and even gensim packages but it seemed to be providing another important insight about the hidden information which I not very trivial on the basis of LDA. Since we intentionally created the topics and we knew the relationships, and thus results were somewhat predictable but with a few surprising elements as well.

The major implementation in Jargon algorithm was to create separate dictionaries for nanoscience and datascience fields, along with the coupled one. This was done again using the nltk tooltik and genism, on emajor function used here was freqDist from nltk which helped in arriving at frequency of a particular word in the dictionary and corpora.

This was then coupled with few mathematical relation based entropy functions and we could finally arrive at the jargon distance and efficiency of communication between the two fields using, the Shannon entropy and Cross entropy functions.

The results of LDA are as follows.

```
Dictionary(140 unique tokens: [u'particular', u'size-depend', u'comput', u
'discoveri', u'enlighten']...)

Top ten words and their frequencies
Topic 1 :
0.126*data + 0.048*learn + 0.032*inform + 0.032*theori + 0.032*model + 0.0
32*gener + 0.032*statist + 0.032*comput + 0.016*within + 0.016*drawn

Topic 2 :
0.007*field + 0.007*scienc + 0.007*nanotechnolog + 0.007*physic + 0.007*ma
teri + 0.007*perform + 0.007*engin + 0.007*broad + 0.007*includ + 0.007*sc
ale

Topic 3 :
0.078*particl + 0.053*bulk + 0.053*nanoparticl + 0.037*properti + 0.029*si
ze + 0.027*unit + 0.027*whole + 0.027*transport + 0.027*accord + 0.027*dia
met

Topic 4 :
0.047*known + 0.040*materi + 0.030*engin + 0.026*chemistri + 0.026*new + 0
.024*year + 0.024*wide + 0.024*mineralog + 0.024*nanosci + 0.024*paradigm

Topic 5 :
0.053*diver + 0.031*molecular + 0.027*equal + 0.027*devic + 0.027*semicond
uctor + 0.027*self-assembl + 0.027*rang + 0.027*organ + 0.027*natur + 0.02
7*nanoscal
```

The topics were heavily composed of terms from one of the fields, like topic 1 , was clearly more influenced by the Data sciences field. And it was no wonder that data was the most frequently used word there. But since both fields are scientific and have Mathematics, analysis and technological applications as a common feature words such as science, engineering and scale are things that could have been common to both which were featured in topic 2. Topics 3-5 seem to be more towards the materials and nanotechnology side. Topic three seems to be again related to measurements and Size.

The Jargon distance algorithm when implemented on this set revealed the relatedness of the fields but not to a very high extent.

Some of the major results are as below.

```
Shannon Entropy(nanoscience) : 4.78415926774
Cross entropy (nanoscience as writer) : 8.84321150814
Efficiency of communication : 0.540997946656
Cultural Hole (Jargon Distance): 0.459002053344
```

For a reader in data science it might take just twice the effort to understand the terms in nanoscience and nanotechnology as for people in nanotechnology itself (Shannon entropy is around 4.8 while cross entropy is 8.8). This might be due to the fact that nanotechnology is a very interdisciplinary field just like data science and moreover the terms in nanotechnology might be very different in topics amongst themselves.

In conclusion I would say, both LDA and jargon distance can be important tools to understand hidden structure not easily understandable via simple reading. They can be used as a highly complementary tools as LDA can demonstrate the common or different topics amongst documents, while the Jargon distance and entropy can define how easy is to navigate through the different topics in a mixture of fields.