1.

The disambiguation part wherein the lexical similarity of an assignee to already registered assignees was determined to determine whether it was present in the NBER dataset and in case of an unsuccessful search the cosine similarity with the k(in this case 5)nearest neighbours was estimated to determine the same assignee mentioned in a different annotation was similar to what I had previously done, hence rang some bells for me. But if there is something like Massachusetts Institute of Technology (MIT) and Manipal Institute of Technology (MIT), I think first the full names instead of the abbreviation is being checked for.

2.

The k-means clustering adopted to group together seemingly similar inventors based on the matches between assignee, date, location, CPC class seems to be a really good idea unlike my previous project wherein I had to go through a cumbersome procedure(a part of it manual) to cluster words with the same spelling but different meaning based on context and paragraph in Cbow modelling and word to vec.

3.

Cosine similarity was used instead of Euclidean distance parameter which I think was more effective since it makes use of the vectorisation and gives the angle and can be easily affected using a short Python snippet (did that in the last project) without using too many python libraries.

4.

The conversion of the initial Patent HTML file to a tsv file and especially the removal of incorrect patents without all 23 attributes was a really user-friendly approach as otherwise trivial errors like a newline character or a missing record or a nan, can cause an unnecessary hassle in an otherwise perfect code and a lot of extra code is required to be run to determine the glitch which takes a lot of time especially when the file is in gigabytes and file splitter is used leading to multiple repetitions of the same code. It is good to rectify and filter data initially.

5.

Also, the character encoding whether Unicode or UTF-8 or Latin-1 is a factor since I have had some problem in the past though Anaconda and Ubuntu work just fine with the proper encoding assigned and some change in the settings in Anaconda and if that’s not enough Latin-1 generally works. The success rates of the matching after the data has been extracted is really high compared to what I had in my last project with a highest match of about ~64% ; which could probably be laid down to the orderly mining of the data.

6.

For the remaining part the graphs are understandable but the table on page 18 wasn’t really succinct.

7.

Lexical novelty as addressed in the last part is pretty interesting and an article I was reading on the lexical novelty of Austin novels like ‘Emma’, ‘Sense and Sensibility’ and so on might provide a reason why studies on lexical novelty are not much accurate owing to the use of features like paragraphs which can vary grossly in size(anyway that is not applicable here).The data shows 1,121,468 patents with at least one new word and 2,816,425 new words in all. So it would probably be a good idea to generate a record of how many times a new word appears in different patents, let’s say a certain word is new and appears in 43,315 patents from the year 1987 with a peak in use from 1988-2001(could be done by plotting occurrence vs year in Plotly).This would provide more insight in to the type of research occurring over a particular time. Moreover, Stream vs new word (over a designated time) can also be plotted to show the most prospective area of research at a particular time. 3,670,047 patents contain no new words or 0 lexical novelty. The topic of research, assignee and dates of these patents could be matched to determine how many of the focused-on topics that had already been introduced previously in 1975-1985 and hence contained terms and topics that were previously registered and not new in 1985-2014.

8.

The work seems pretty interesting and many issues can still be addressed in the lexical novelty part, like the number of times a new word has ben used by a certain inventor and his co-inventors, and by inventors with no link to him, if the assignee organisation is the same, whether the country is the same if the assignee org is different to throw light upon research trends in an organisation or region at some time. The number of new words introduced per region can also be plotted as a function of the region to gauge the growth or decline of research activities in an area over time. And such other interesting parameters can be determined once the data is mined and cleaned.

Exploring the applications of Random Forest for patent classification could propose to be an interesting topic.