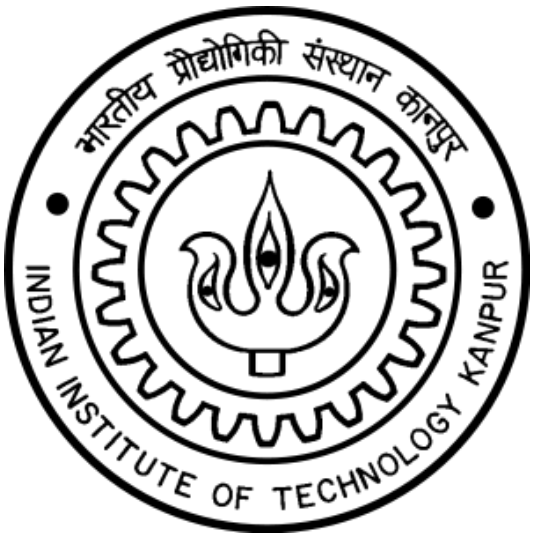


Diachronic Word Sense Change Identification

Ankit Singh(12128)
T Raghuveer(12762)

Mentor - Prof. Amitabha Mukherjee
Indian Institute of Technology, Kanpur Undergraduate Course Project

Abstract

Language has continuously changed over time. New senses for some words took birth while some old senses demised. For an example the word 'instant' which referred to the current month about a century ago but is now used to refer to a very small brief time; 'economy' referred to management of resources in the past and is now being referred to state of country in terms of productions. Diachronic word meaning change identification has been done by making vectors of words using Word2Vec[5] over different time epochs; aligning the vectors across the epochs and finally used cosine similarity to identify the words.

Contents

1 Introduction & Motivation

Diachronic means dealing with phenomena (as of language or culture) as they occur or change over a period of time. Language has continuously changed over time. New senses for some words took birth while some old senses demised. For an example the word 'Artificial' had a positive sense in the previous centuries but now has a negative sense associated to it. Some other examples include 'gay' which meant a noble person in the previous century but is now used to refer to a sexual orientation; 'sick' referred to illness in the past and is now being referred to as something crazy or cool.

It has been theorized that polysemy in language is simply a transitory phase of word evolution where a new sense evolves with time and competes with existing senses. This time sense disambiguation is also highly instrumental in culturonomics; to analyze the changes in human culture and historical phenomena by evaluating usage of various words. Understanding the changes in meaning and usage of words is highly important for people working with historical texts, such as librarians, historians and linguists. It is also helpful to lexicographers and design engineers in a variety of NLP/IR tasks.

Given the availability of Diachronic datasets and the computational ability, we in this current work, aim to look at evolution of semantics over time; identify and report words where change of sense has occurred.

2 Related Work

Sunny Mitra et al. [4] constructed distributional thesauri of historical datasets and cluster each of them separately to obtain word-centric sense clusters for different time points and compare these sense clusters of two different time points to find the birth of a new sense; split of an older sense has into different sense; formation of a newer sense and dying of the senses.

Adam Jatowt et al. [2] diachronic word sense change identification disambiguate words by measuring cosine similarity of co-occurrence matrix vectors created using Google books 5-grams model.

Derry Tanti Wijaya et al [7] identified clusters of topics surrounding the entity over time using Topics-Over-Time (TOT) and k-means clustering. They conducted analysis on Google Books N-gram datasets and showed how clustering words that co-occur with an entity of interest in 5-grams can shed some lights to the nature of change and identify the period which the change occurred.

3 Problem Statement

The task at hand is to devise an unsupervised approach to determine semantic change, i.e. transition in sense of words over time based on extensive analysis of diachronic text data and to visually present the word sense changes identified in over a time frame.

4 Dataset Collection

4.1 English

British parliamentary debates from 1803 were been scanned from hard copy sources and digitized as Hansard Archive[3]. The debate corpus was extracted from archive as ZIP files with total of around 1500 zip files. Each zip file was extracted and each zip file contained debates of a particular year in XML format. The extracted XML files contained debates over the time period 1892 to 2005. These XML files were pre-processed to extract year of the debate and text of correspondence. The whole corpus was divided into 12 epochs with century tags with them.

4.2 Hindi

Story, fictions, novels, plays, essays and poems were extracted from Hindi Samay [6]. Multiple HTML files were processed, with multiple level scraping(within each html) being done to extract the time range of the document from author details

5 Methodology

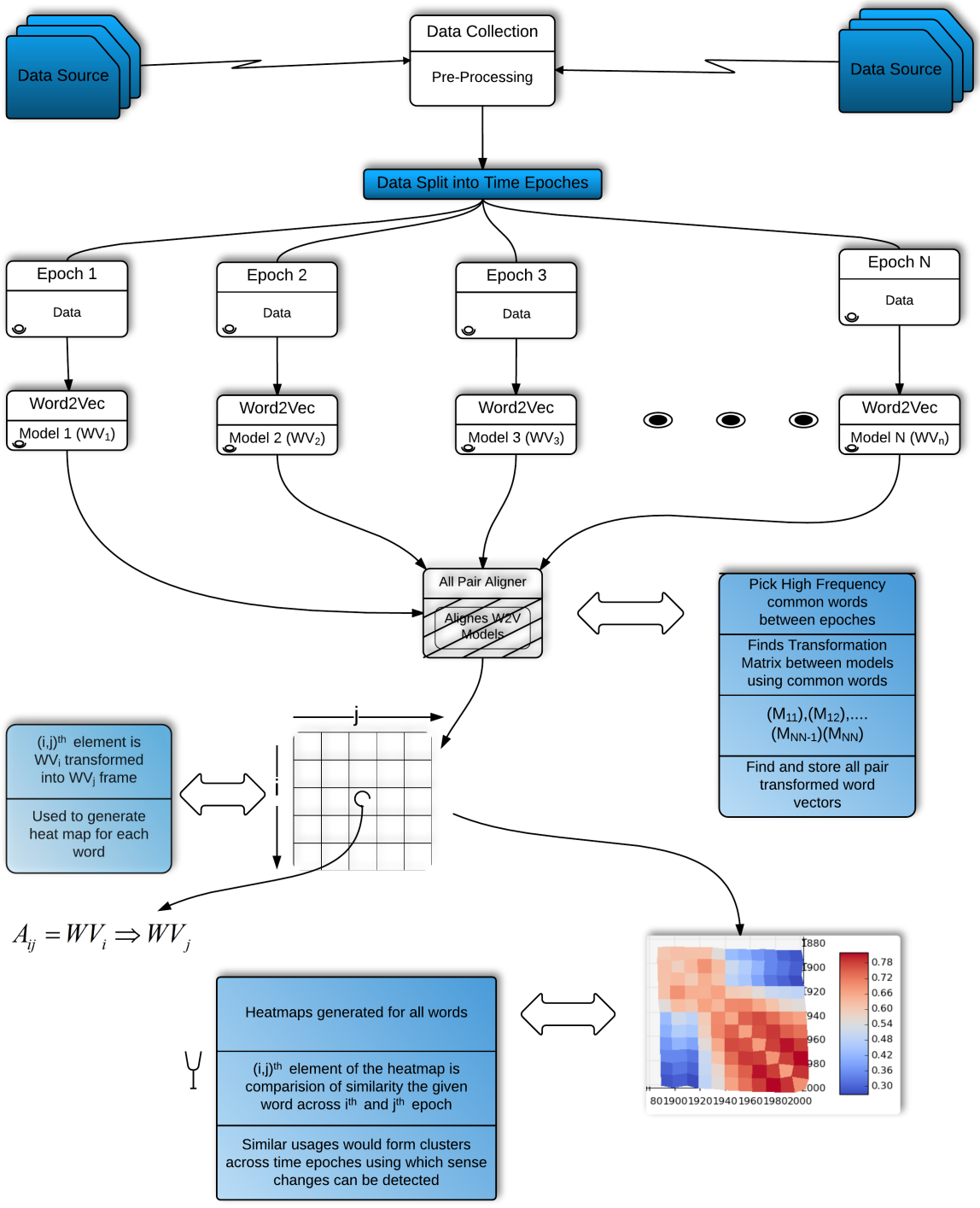


Figure 1: Methodology

5.1 Procedure

This work is based on Shashwat Chandra's M.Tech thesis of 'Aligned Word Vector Spaces and Document Vectors' [1]. Word2Vec is trained on all epochs with dimension of 200 and minimum word count greater than 3. Now we have 12 word2vec trained model WV_1, WV_2,..., WV_12; for each of the centuries from 1890 to 2000. Now we transform each of the word2vec model into the other word2vec model. Here we have used the assumption that one word2vec model can be linearly transformed to another word2vec model. Using this assumption, we calculate transformation matrix for all the model pairs using the following formula:

$$W_i = M.W_j + b$$

$$W_i: Transformed\ word\ vector\ matrix$$
$$M_{ij}\ Transformation\ matrix$$
$$W_j: Word\ vector\ from\ initial\ set\ matrix$$
$$b: Bias\ term\ for\ translation$$

Converting the above formula in matrix formation we get:

$$\begin{bmatrix} W_i \\ 1...1 \end{bmatrix} = \left[\begin{array}{c|c} M_{ij} & b \\ 0...0 & 1 \end{array}\right] . \begin{bmatrix} W_j \\ 1...1 \end{bmatrix}$$

$$[M_{ij}|b] = W_j.(W_i)^+;$$
$$W^+ = W^T(WW^T)^{-1}$$

In estimation of M_{ij} and b, we pick 600 most frequent words across all the epochs whose meaning has been not changed over the time(few hand-picked and few based on the frequency of their occurrence); concatenate their vectors to form W_i and W_j matrix of 201*600200 being the dimension o the word vector and one extra for the bias and use the Least Square solution to find an estimate of M_{ij} and b.Here the matrix M_{ij} is 200*200 matrix and b is 200*1 matrix and this transformation matrix is calculated for all epoch pairs forming total of $({}^NC_2)$ pairs(as transforming i to j or j to i same).
Using this transformation matrices, we calculate the transformed dictionaries A_{ij} for all the pairs containing the transformed word vector for each of the words where A_{ij} element corresponds to transformed dictionary of words from i^{th} epoch into j^{th} epoch. A_{ij} is calculate as follows:

$$A_{ij} = [M_{ij}|b].W_i$$

$$[M_{jj}|b] = I$$

Now, given a word we construct a heat-map using the transformed models as follows.

$$HeatMap(i, j|word) = cos\_sim(A_{ij}[word], A_{jj}[word])$$

In a heat-map, similar senses of a word are clustered over Epoch axes and hence change in sense of a word can be detected by seeing the cluster formation. We only consider words with maximum deviation was greater than 0.6 and standard deviation greater than 0.3; plot them and identify sense changes.

6 Results

Some words were found to have two different similarity clusters in their heat map indicating change in their sense over time. The intersection point of these two clusters indicates the epoch where the change in sense appeared.

6.1 Positive examples

Most of the words with two clusters in their heat map were found to be used in different senses. All examples are provided in the Appendix

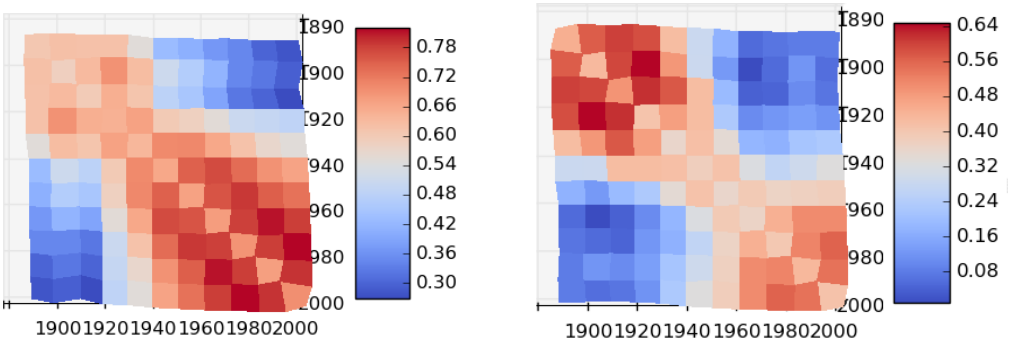
Words in Fig 2, Fig 3 i.e. 'economy', 'instant' have changed/added senses in the corpus in the two clusters formed as shown.
Instant – month to brief instant
'Economy' – management to financial market
Here are some use of word 'economy' and 'instant' used in our dataset:

'Economy'

```
1900:  management
-we are told we can look with no hope for economy in the permanent
   services of the country
-the promises of economy have resulted in an increase of
-i do not say a niggardly regard for economy as such
-friend's political economy is sound only in parts

1990s: - finance related
-partly because we have operated an open economy
-the economy and britain's competitiveness in world markets
-instead of shifting the weight of the economy from domestic demand
   to export-led demand
-the queen's speech referred to improving the economy through the
   framework of the anglo-  irish agreement
```

'Instant'

```
1920: month related time
-17th instant
-20th instant

1990s: - brief time interval
-would have unselfishly given his instant support.
```



(a) Heatmap for 'economy'

(b) Heatmap of 'instant'

6.2 False Positives

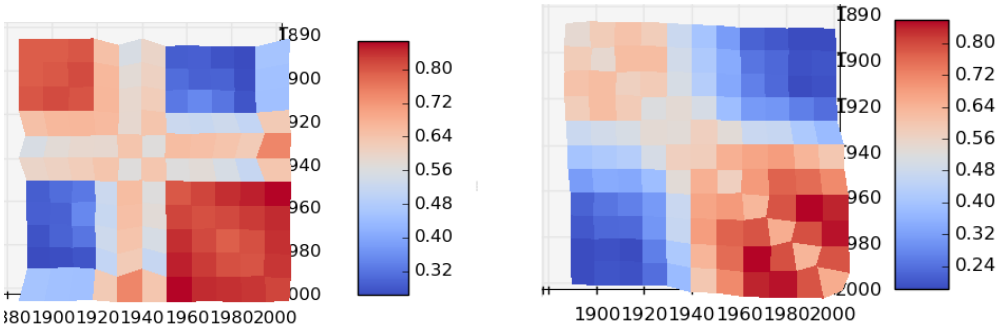
Some words eventhough they had 2 clusters formed in the heatmaps, were not exactly found to have changed sense
The word 'east' and 'confirm' correspond to *False positive* examples, because of varied contexts that appeared in the corpus.
'East' - had very varied contexts

```
1920
- east end of london
- in the county of east sussex
- east african colonies
- east indian association of british guiana

1980
- with the university of east anglia
- member for east surrey
- member for east lothian
```

'Confirm' - again had varied contexts

```
1920 :
- parliament a bill to confirm the order
- an act to confirm a provisional order
- confirm me in the statement that my hon
```

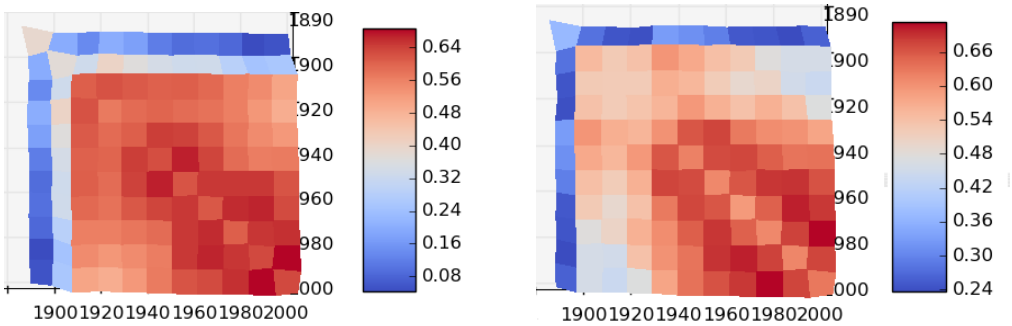
```
- only confirm what the british government are prepared to give us

1980
- will he confirm that
 - scientific appraisal to confirm their safety
- friend confirm that there is nothing to stop local education
    authorities
- will the secretary of state confirm that there will be no
    announcement tomorrow of increased capital funding of ctcs
```



(a) Heatmap for 'east' (b) Heatmap of 'confirm'

6.3 Negative examples

The 'deflected' which kept same sense over epochs. The row and column corresponding to 1890 epochs shows low similarity with other epochs because the word 'deflected' occurred with low frequency. Also the word 'buses' had not changed its sense over a huge time span



(a) Heatmap for 'deflected' (b) Heatmap of 'buses'

7 Conclusion

We propose and implement an unsupervised technique for identification of words with changes senses over a time axis corpus. The most easily avail-

able corpus google 5 gram would not work with our technique of training Word2Vec. Hence another historical corpus was extracted out.

We train Word2Vec for each of the epoches extracted and find transformations between all pairs of epoch vector models. We then use the concept of heatmaps to find words with changes meanings.

8 Future Work

- Handle polysemy words present within one epoch and appropriately train word2vec separately for different words within an time epoch; propose birth and death of words; and merge and split of the senses for words.
- Use of Distributed thesaurus to train Word2Vec over Google 5-gram model and see the results as all the previous works has been done on the Google n-gram model.

9 Acknowledgement

The work has been done as a part of the course CS671A: Natural Language Processing. We would like to thank Prof. Amitabha Mukerjee and the TAs for their useful insights and continuous support and guidance throughout the project.

10 Source Code

The source code of the work and the dataset used is available at the :
`https://goo.gl/dW7ZKP`

11 Appendix

11.1 Positive examples

The following were some words which had changed their meanings(alphabetic order). Varies usages shown from the center epoch of clusters fromed in heatmaps.

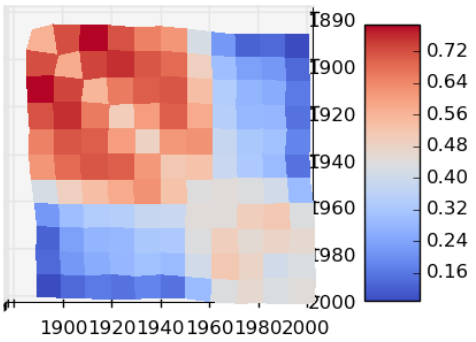
In most cases the overlapp is not completely zero as can be seen from the blue region of the heatmap. The most common used and their meaning sense in different epoches are shown below

- Admiralty

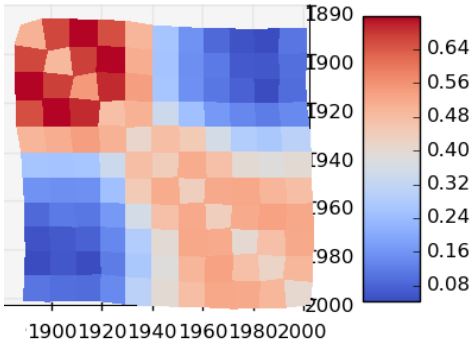
```
1890 : court of navy
- i beg to ask the first lord of the admiralty if he can now
    state whether a training ship for boys of the royal navy is
    to be stationed in harwich harbour

1980 : general navy related
- admiralty nuclear research
- admiralty research establishment
- admiralty pier bills
- lord of the admiralty
- admiralty surface weapons establishment
- admiralty floating dock
```

- affirmative

```
1890 : diff usage
- the answer to the first paragraph is in the affirmative
- hearing that some had received a reply in the affirmative

1980: positive, confirmed
- affirmative orders
- affirmative action constitutes a violation of the
    prohibition on discrimination
- affirmative resolution will be debatable
have not an affirmative judgment in our favour
```



(a) Heatmap for 'admiralty'



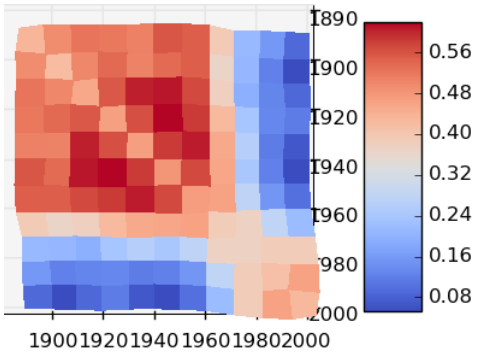
(b) Heatmap of 'affirmative'

- breast

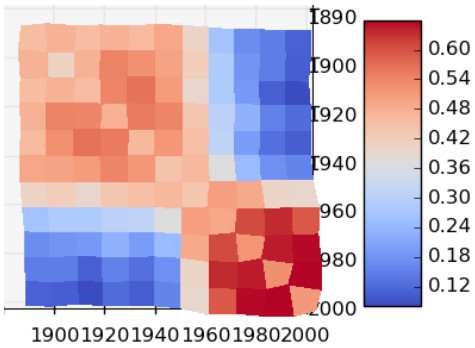
```
1920 : chest/heart
- beating his breast
- the world will never know the struggle which went on in his
    breast at that time
- the secrets in another's breast
- share the deep feelings which stir the breast of the hon

1980 : contemporary meaning of breast
- evaluation of the early detection of breast cancer
- salts for breast surgery
- evaluation of the early detection of breast cancer
```

- challenges

```
1910 - question/
- any other country a system which challenges the authority of
    the state
- there have been two challenges given to the government
- whenever anybody challenges their conviction
- challenges the statement made by not merely the two actuaries

1980: problems
- industrialists will be faced with considerable challenges
- in the environmental decade we shall take hold of all the
    hard-fought gains of the 1980s - - and hit the challenges
    of the earth
- with all the competitive challenges
- moment the external economic challenges that face western
    europe
- national health service always faces fresh challenges
```



(a) Heatmap for 'breast'



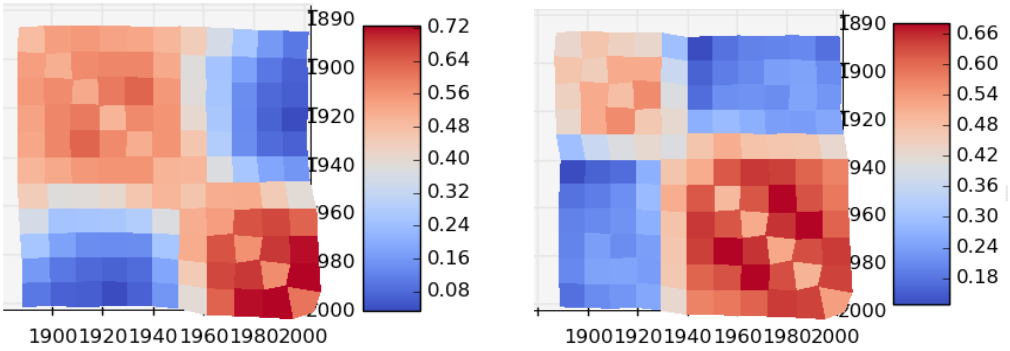
(b) Heatmap of 'challenges'

- commissioned

```
1920 : assigned
- flying was necessarily done by the commissioned ranks
- commissioned medical officers with over three years' war
    service compulsorily retained in
- who when commissioned in the air force were serving as
    warrant officers

 1980 : conducted
- commissioned research and development would depend on a
    number of factors
- the scottish salmon growers association have commissioned
    the institute of aquaculture at the university
- we have just received the preliminary results of research
    commissioned by the department of health from dr
- a survey of sight test numbers will be commissioned later
    this year
```

- controls

```
1910 : power/ authority
- controls which existed under certain governing conditions to
    the situation which has arisen
- but in one of the parts of the country the pipe-line of some
    hundreds of miles has been at -the mercy of the tribes whom
    no power controls
- government still controls the stock of hay in this country

1980 : limit
- exhaust emission controls which we fix to our vehicles
- credit controls do not and could not work
- what quality controls are used by (i)
- responsible for welfare controls
```



(a) Heatmap for 'commissioned'

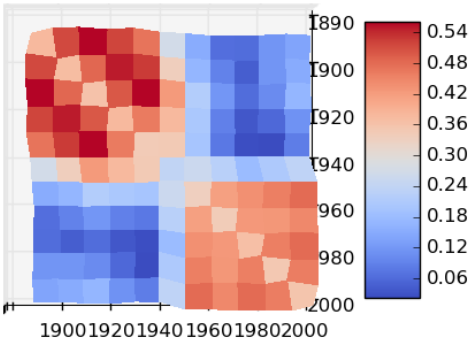
(b) Heatmap of 'controls'

- dispatch

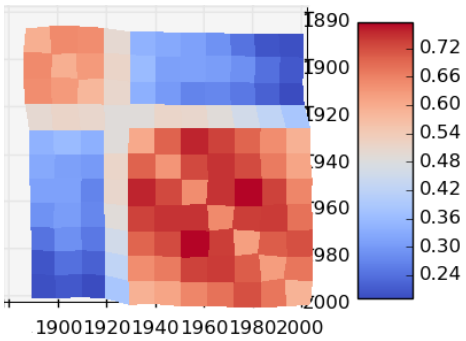
```
1920 : deliver
- with the dispatch or the delivery of goods and in any
    clerical capacity
- railway has not given quicker dispatch
- the dispatch sent in reply to an inquiry
- whether it has been decided to dispatch to china any
    fairey-napier aeroplanes;

1980 : conferance table
- office question time appearance at the dispatch box
- i repeat the promise made from this dispatch box on many
    occasions
```

- ease

```
1920 : case
- in any ease
- in ease of any dispute as to the existence or scope of any
    record as aforesaid
- there may be hundreds killed in such a ease
- it was not a ease of its having been forced through with the
    aid of the government whips

1980 : loosen
- curtail trade union powers will not ease industrial problems
    in selby
- ease the transition from rates to community charge
- to build bigger and better roads to ease congestion is not
    the answer
- a harbour launch which will ease the transfer of passengers
    and baggage between the rms st
```



(a) Heatmap for 'dispatch'



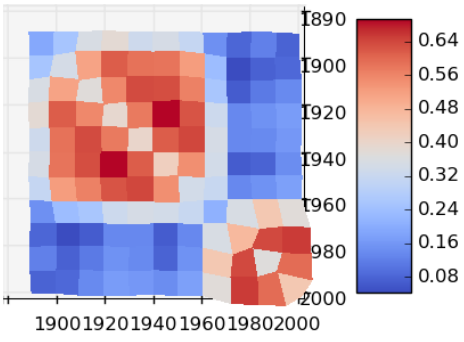
(b) Heatmap of 'ease'

- environment

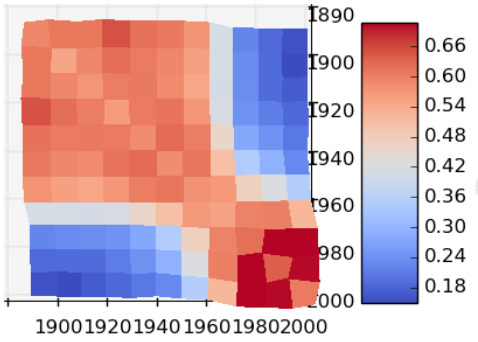
```
1930 - background
- they are divorced from their families and are carried into a
    new environment among strangers
- not comparable with ours because of different environment
    and circumstances
- too often the result of the bad environment under which they
    live
- that shows the tremendous effect of the slum environment

1990 - nature around us
- the environment and regions is highway authority and the
    highways agency acts on his behalf
- under section 59 of the environmental protection act 1990
    the environment agency and local  authorities may remove
    flytipped waste
```

- remit

```
1910
- to remit the fee if necessary
- nd i am clearly in mo position to remit taxation
- is now serving his third sentence in maidstone prison; and
    whether he will forthwith remit  this man's sentence

1980
- the cabinet committee on rural affairs already has the remit
    of co-ordinating the   government's policies
- it is also part of the remit of many of the department's
    advisory committees to advise on -  new scientific
    discoveries and to present recommendations for further
    research
- the rdas have a remit to deal with and report on such matters
```



(a) Heatmap for 'environment'



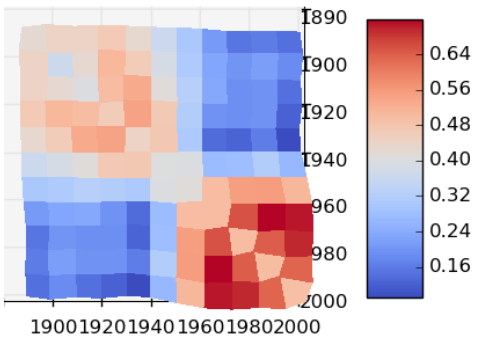
(b) Heatmap of 'remit'

- spell

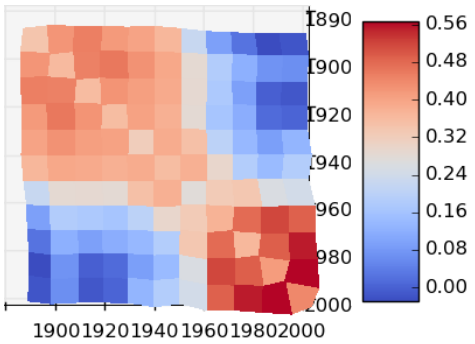
```
1920 - time interval
- undoubtedly the prolonged spell of industrial depression
- if i shall not break the spell by stating the facts
- during the cold spell a year ago the quota was changed
    several times in one month
- you may get a very high figure during some temporary spell
    of cold weather

1980 - spelling of word
- i feel that i must spell out to him what is involved
- in which he tried to spell out the nature of citizenship
- we have to spell out that the flooding is almost certainly a
    result of climate change
```

- tailor

```
1920 : cloth tailor
- there is the example of the tailor who
- i feel that while the argument of the baker and of the small
    tailor has
- the tinker and the tailor to go into trade with a small
    capital and become rich in a comparatively small years

1990 : shape up
- advisers discuss with the individual their circumstances and
    skills and tailor the help that will best equip them to
    enter the labour market
- we need to tailor arrangements to suit the particular
    circumstances of the young person concerned
- he local government bill provides a tailor made framework
    for executive arrangements including cabinets.
```



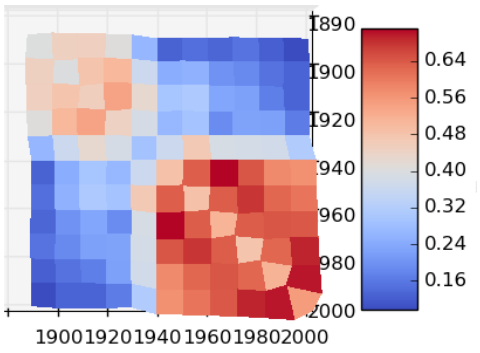
(a) Heatmap for 'spell'



(b) Heatmap of 'tailor'

- target

```
1920 : target as in war
the war department standard pattern "hythe" iron target
reid was offered 0 in full discharge of all claims against his
    majesty's government and in full recognition of his
    services rendered in connection with target frames
 baronet has missed his target


 1980 : goal/aim
gentleman is arguing for an immediate commitment to a target
    for emission reductions
e is right to suggest that we are arguing for the setting of a
    target and a time scale
the normal target is four regular officers on watch
```



(a) Heatmap for 'target'

11.2 Negative examples

All other heatmaps of words which did not give two clusters were negative examples. There were over 20000 words. Among these only the prominent ones with high variance i.e. variance in the upper triangle in heatmaps over 0.3 have been uploaded on to the link of the code. All heatmaps with less than variance of 0.3 were sure to have been in a single sense all through the epoches

References

[1] Shashwat Chandra. "Aligned Word Vector Spaces and Document Vectors". In: (2015). DOI: http://172.28.64.70:8080/jspui/handle/123456789/15193.

[2] Adam Jatowt and Kevin Duh. "A Framework for Analyzing Semantic Change of Words across Time". In: (2014). DOI: http://cl.naist.jp/~kevinduh/papers/jatowt14change.pdf.

[3] UK Parliament. "Hansard archive (digitised debates from 1803)". In: (). DOI: http://www.hansard-archive.parliament.uk/.

[4] Martin Riedl Chris Biemann Animesh Mukherjee Pawan Goyal Sunny Mitra Ritwik Mitra. "Thats sick dude :Automatic identification of word sense change across different timescales". In: (2014). DOI: https://www.aclweb.org/anthology/P/P14/P14-1096.pdf.

[5] Greg Corrado Jeffrey Dean; Tomas Mikolov Kai Chen. "Efficient Estimation of Word Representations in Vector Space." In: (2013). DOI: http://arxiv.org/pdf/1301.3781.pdf.

[6] Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya. "Hindi Samay". In: (). DOI: http://www.hindisamay.com/.

[7] Derry Tanti Wijaya and Reyyan Yeniterzi. "Understanding Semantic Change of Words Over Centuries". In: (2011). DOI: http://rtw.ml.cmu.edu/papers/wijaya-detect11.pdf.