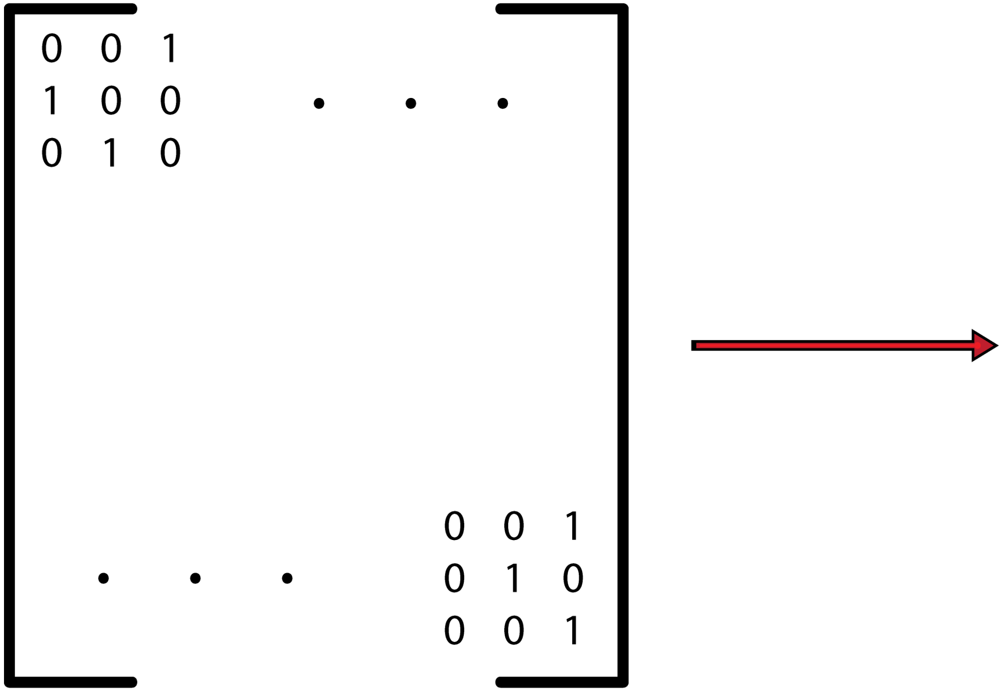
Estimating the Optimal Number of Latent Concepts in Source Code Analysis

Scott Grant James R. Cordy

Queen's University
Kingston, Canada



Topic Token List	Matching Methods
lorem ipsum dolor sit amet consectetur adipisicing elit	/path/to/file/source.c, method1() /path/to/file/source.c, method2() /path/to/file/source.c, method3()
sed do eiusmod tempor incididunt ut labore et dolore magna aliqua	/path/to/file/source.c, method4() /path/to/file/source.c, method5() /path/to/file/source.c, method6()
ut enim ad minim veniam quis nostrud exercitation	/path/to/file/source.c, method7() /path/to/file/source.c, method8() /path/to/file/source.c, method9()

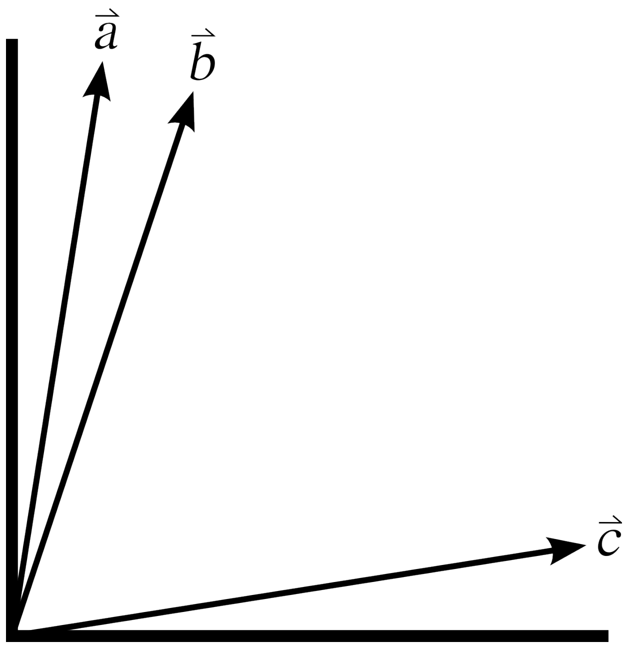
Question: In order to gain the most benefit from latent models like Latent Dirichlet Allocation, how many *latent concepts* should we assume exist in the source code?

Motivation: Choosing a good latent concept count improves the results obtained with the model.

What's the goal?

The goal of this research is to identify a method for obtaining an estimate of the optimal number of latent topics for source code analysis.

In order to do this, we propose using known metrics about the data set as estimates on the document-by-document relationships.



Heuristic 1: A method is likely to be conceptually related to its nearest neighbours in the vector space representation of the source code methods.

Heuristic 2:Two methods are likely to be conceptually related if they are found in the same file or folder.

/my_code/foo/bar/file_a.c

/my_code/foo/bar/file_b.c

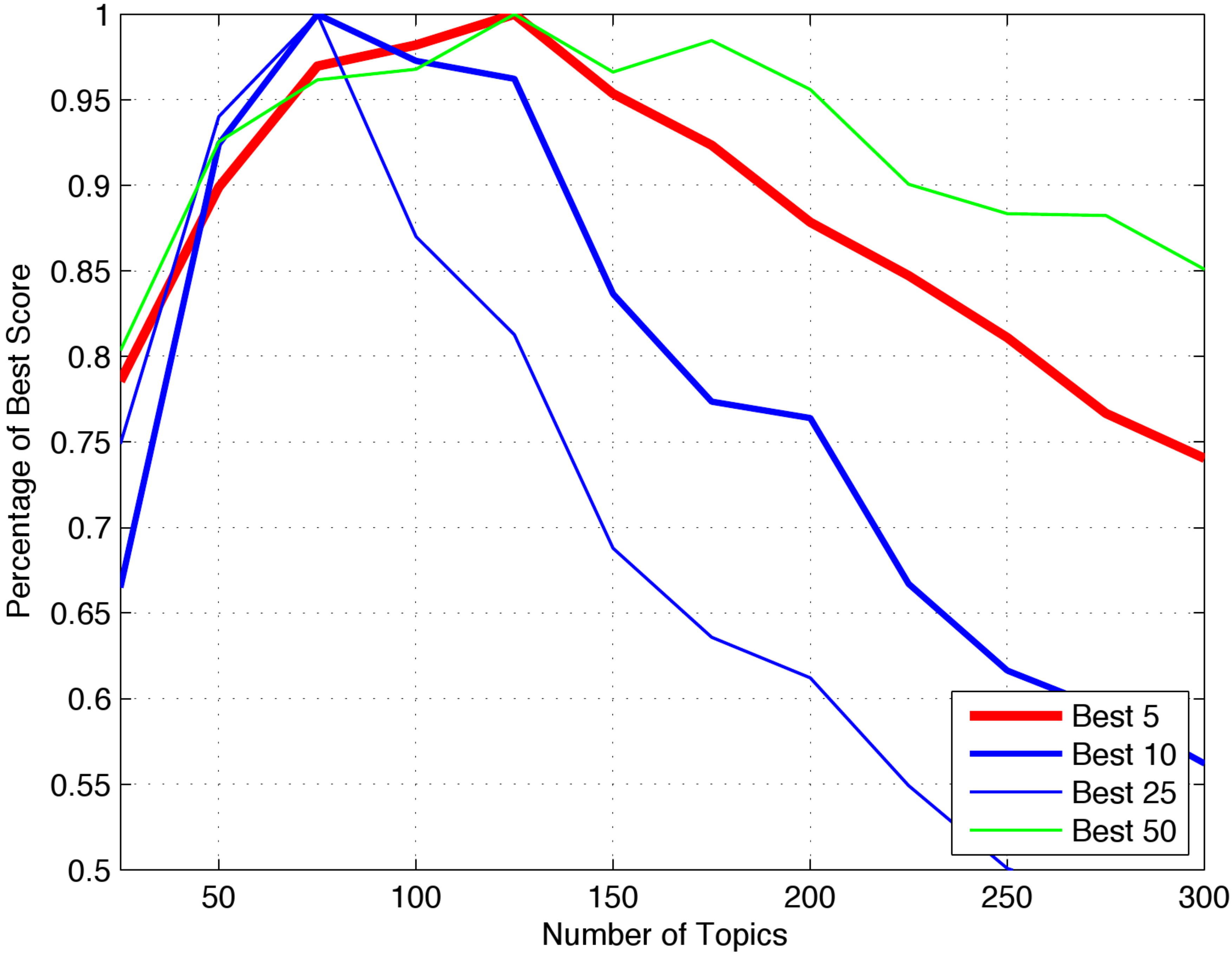
...

/my_code/bash/file_c.c

Overall Nearest Neighbour Score

Generate an LDA model Mk, with k topics

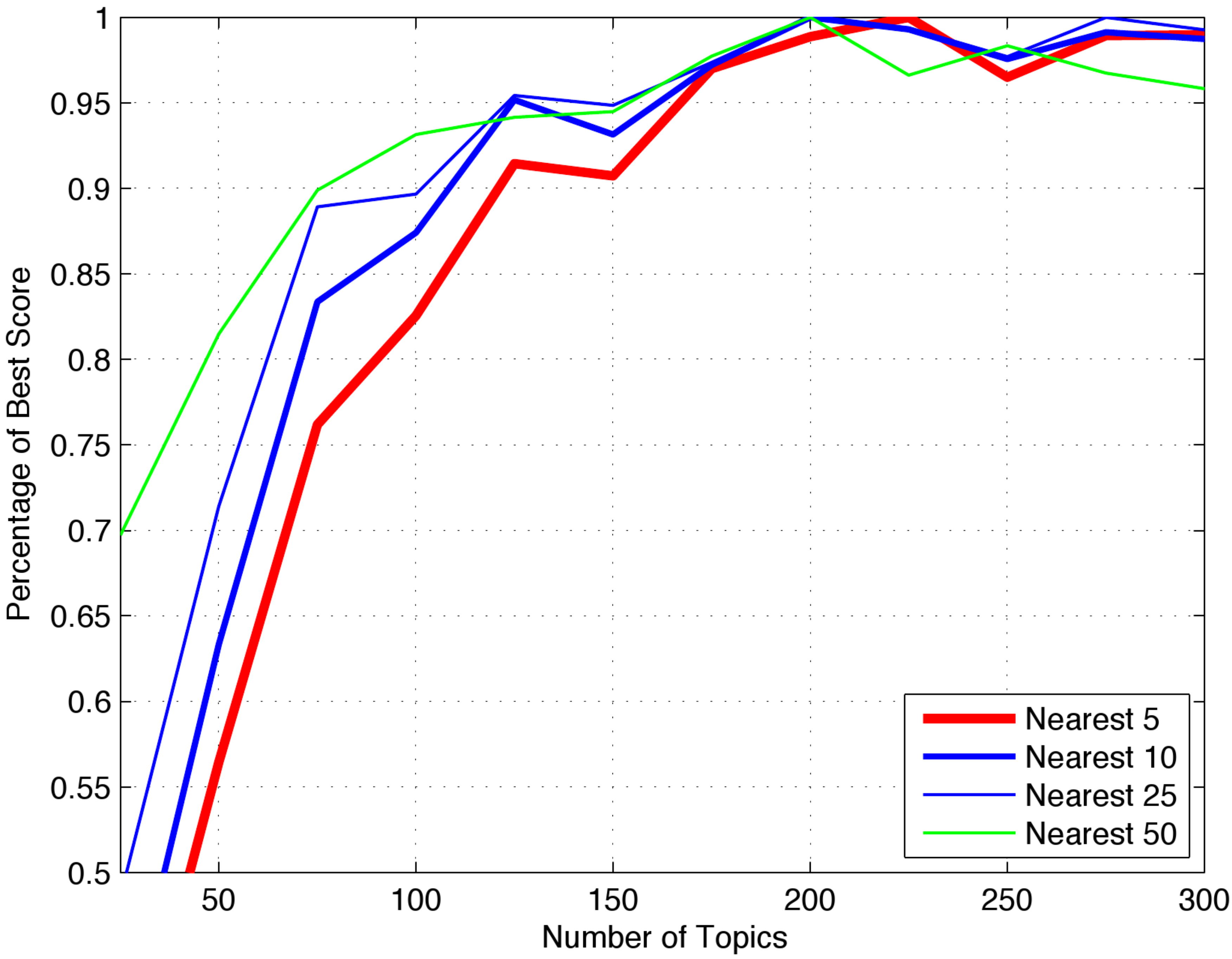
For each document in Mk, get the m nearest neighbours by cosine distance, and calculate the number of documents satisfying Heuristic 2. The average over all documents is the overall nearest neighbour score.


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Percentage of Best Score
50
100
150
200
250
300
Number of Topics
Best 5
Best 10
Best 25
Best 50

Overall Topic Score

Generate an LDA model Mk, with k topics

For each topic in Mk, get the m documents that have the highest probability of belonging to this topic, and calculate the number of documents satisfying Heuristic 2. The average over all documents is the overall topic score.


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Percentage of Best Score
50
100
150
200
250
300
Number of Topics
Nearest 5
Nearest 10
Nearest 25
Nearest 50

Further Research

Code Proximity is only one available metric, there are many more available.

We have continued the research using clone detection as an oracle for conceptual relationships, and continue to explore what these relationships actually tell us about the similarity between code fragments.

Conclusions

Using existing knowledge about source code can help tune parameters in latent models.

The number of topics used with source code in a latent model must be carefully considered.

Controversy

Software engineers don't care about concept location.