# Data Visualisation and Data-Driven decision making

# The Creation of Multi-Domain Lexicons

By

Viktor Torp Thomsen
vikt@itu.dk

4th Semester, Spring 2019
Course code, BSDVDDM1KU

IT UNIVERSITY OF COPENHAGEN

# Contents

# 1 Introduction

Trying to understanding the concepts of a new machine learning model can often be a confusing journey. The aim of this project is to explore how data visualizations can be utilized as a didactic tool to help ease this journey. The purpose these visualizations are to describe how, multi domain lexicons are created, to an audience with some prior knowledge in natural language processing(NLP) and machine learning. These lexicons are used to find and map concepts from one domain to another domain, e.g "written" in "book" reviews would be "composed" in "dvd" reviews [3]. The data used in this project originates from amazon reviews and has already been prepossessed to a bag of word representation only containing lower cased letters. These reviews are from 4 different domains "books", "dvd", "electronics" and "kitchen". The raw data and the prepossessed is available at `https://github.com/jbarnesspain/domain_blse`[1].

# 2 Identifying concept words

In order to find concept words the tf-idf score is used. Tf-idf a numerical measure which indicates how important a word is to a domain in a collection of other domain[5]. To visualize which words tf-idf targets, and how these words scores in other domains, only the five highest scoring words from each domain was select. Since the purpose of this visualization (fig. 1) is to help the audience understand how the words tf-idf scores changes, from one domain to another, it was created as an interactive dashboard so that the audience could explore and compare different domains on their own [4, p.10-36]. The dashboard can be seen at `https://public.tableau.com/profile/viktor.torp.thomsen#!/vizhome/Toptfidfscores/TF-IDFDashboard`.

# 3 What should a concept word be mapped to and how to find it.

In NLP, words are commonly represented as vector with the advantaged that one can use vector algebra on the words, e.g. $\vec{dvd} - \vec{composed} + \vec{books} = \vec{written}$ (this type of vector will not always be an actual word vector, so the word vector with the highest cosine similarity is chosen as the new word)[3]. But, imagining how vectors moves and interacts in some high dimensional space can be a rather difficult task, and so it is often beneficial to use some low dimensional representation of the same vectors to visualize these concepts. Therefore, in order to visualize how to find these lexicon pairs I have chosen to use three plots highlighting the different steps in the process (figs. 2 to 4).

# 4 The final lexicon pairs

The final visualizations purpose is to show the actual lexicons, and once again explain the relationship between the new and the old concept word, e.g. what "dvd" is to "actors" is what "books" is to "writers". Furthermore, this visualizing also displays the cosine similarity with these new words in order to show how "confident" we are in these lexicon [2, p.112-116]. In order to make the result more clear in this visualization (fig. 5) only the result from the domain "dvd" are shown.

# 5 Figures

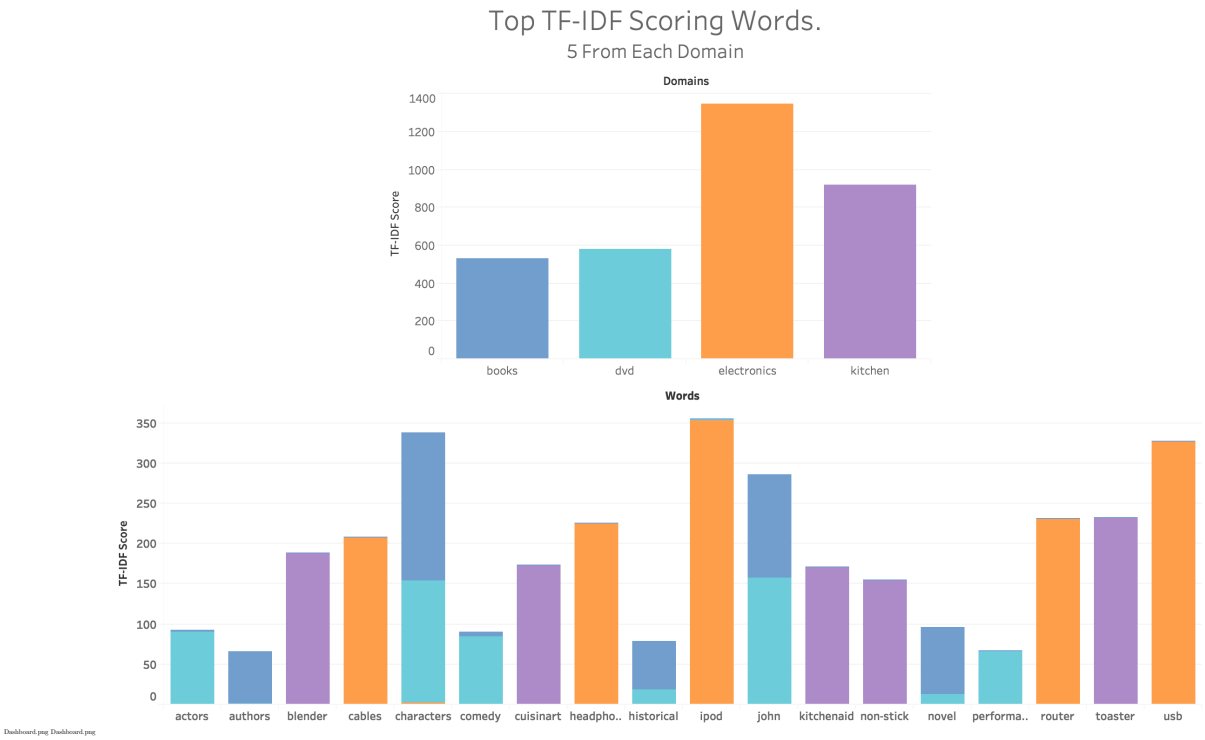


Figure 1: Static image of a interactive dashboard displaying the top 5 scoring tf-idf words for each domain. The dashboard is availeble at
`https://public.tableau.com/profile/viktor.torp.thomsen#!/vizhome/Toptfidfscores/TF-IDFDashboard`

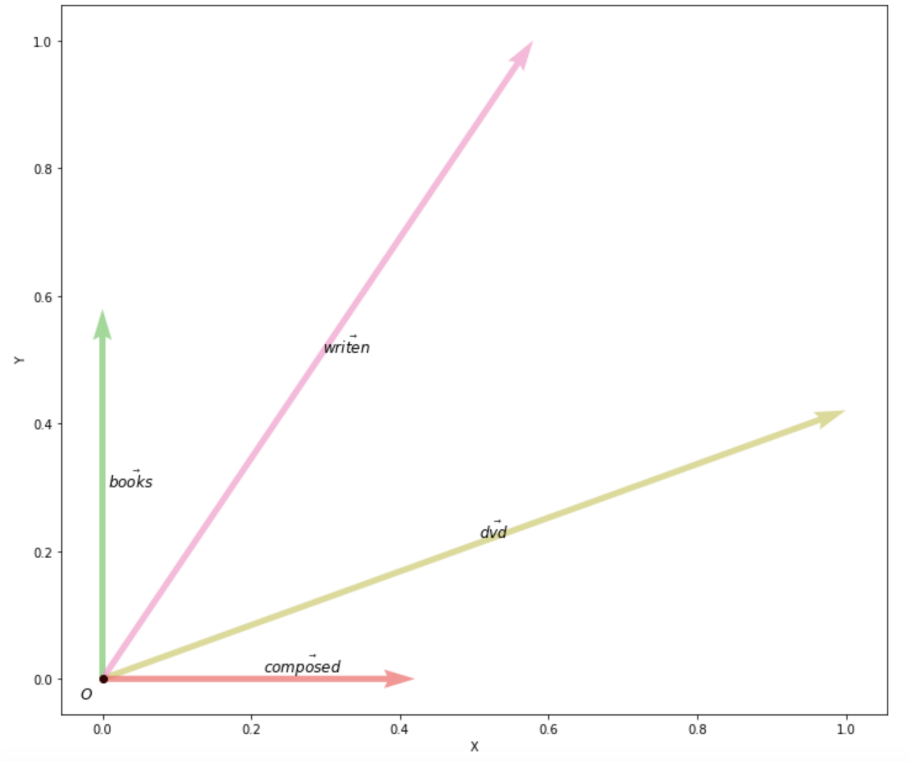


Figure 2: Figure displaying the word vectors for "books", "written", "dvd" and "composed".

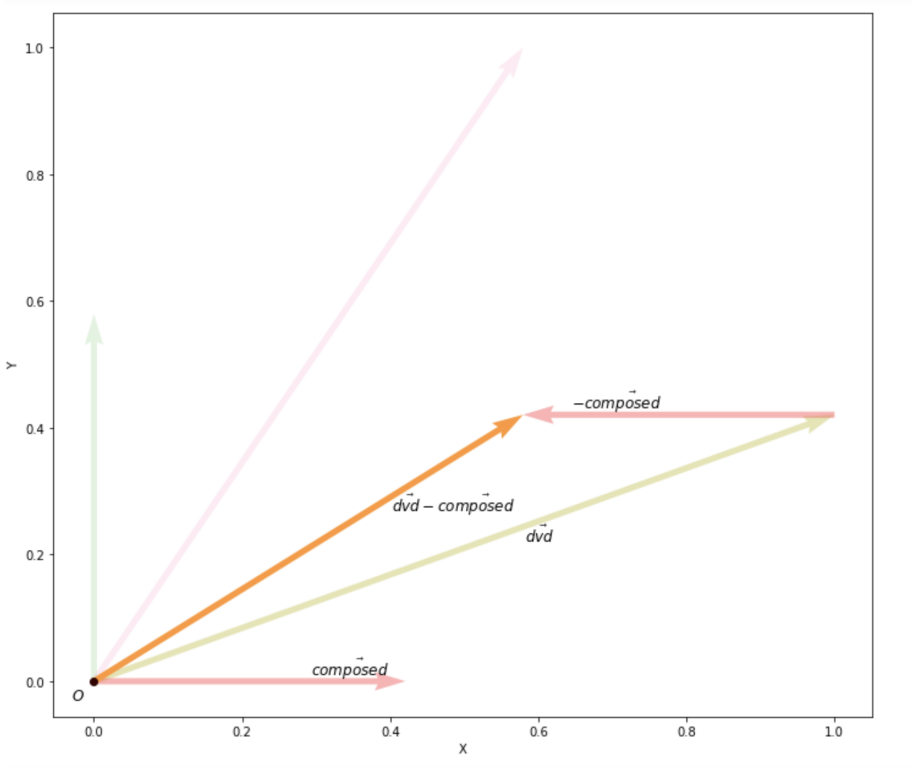


Figure 3: Figure displaying how the difference vector $\vec{dvd} - \vec{composed}$ is created, is by subtracting the concept word "composed"'s vector from the source domain "dvd"'s vector.

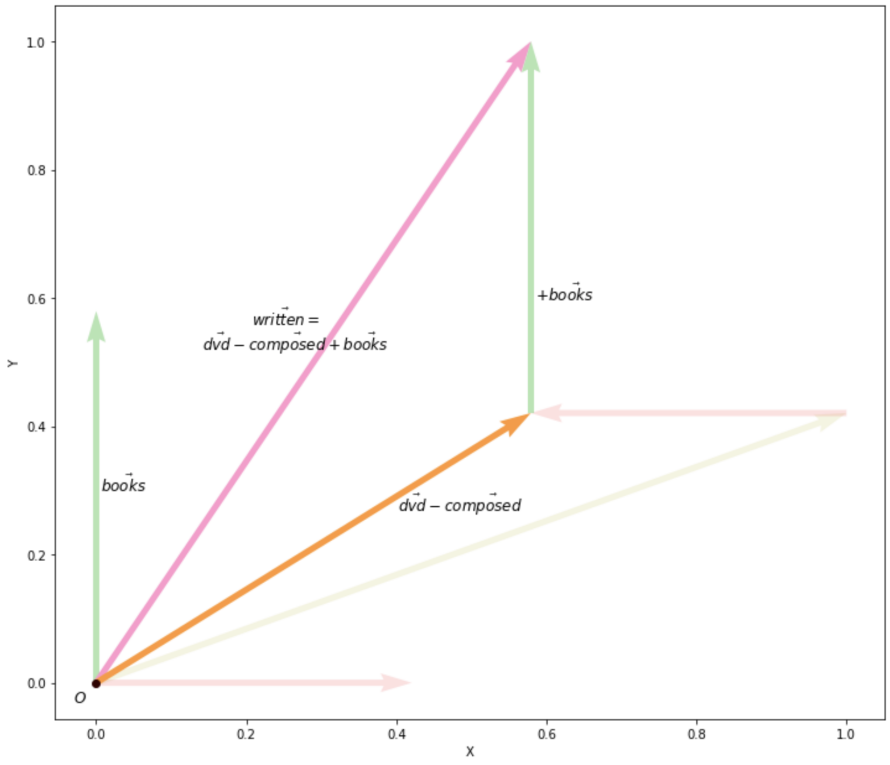


Figure 4: This figure visualizes how the vector $\vec{written}$ is the same as the difference vector $\vec{dvd} - \vec{composed}$ (from fig. 3) plus the target domain vector $\vec{books}$, i.e. $\vec{written} = \vec{dvd} - \vec{composed} + \vec{books}$.

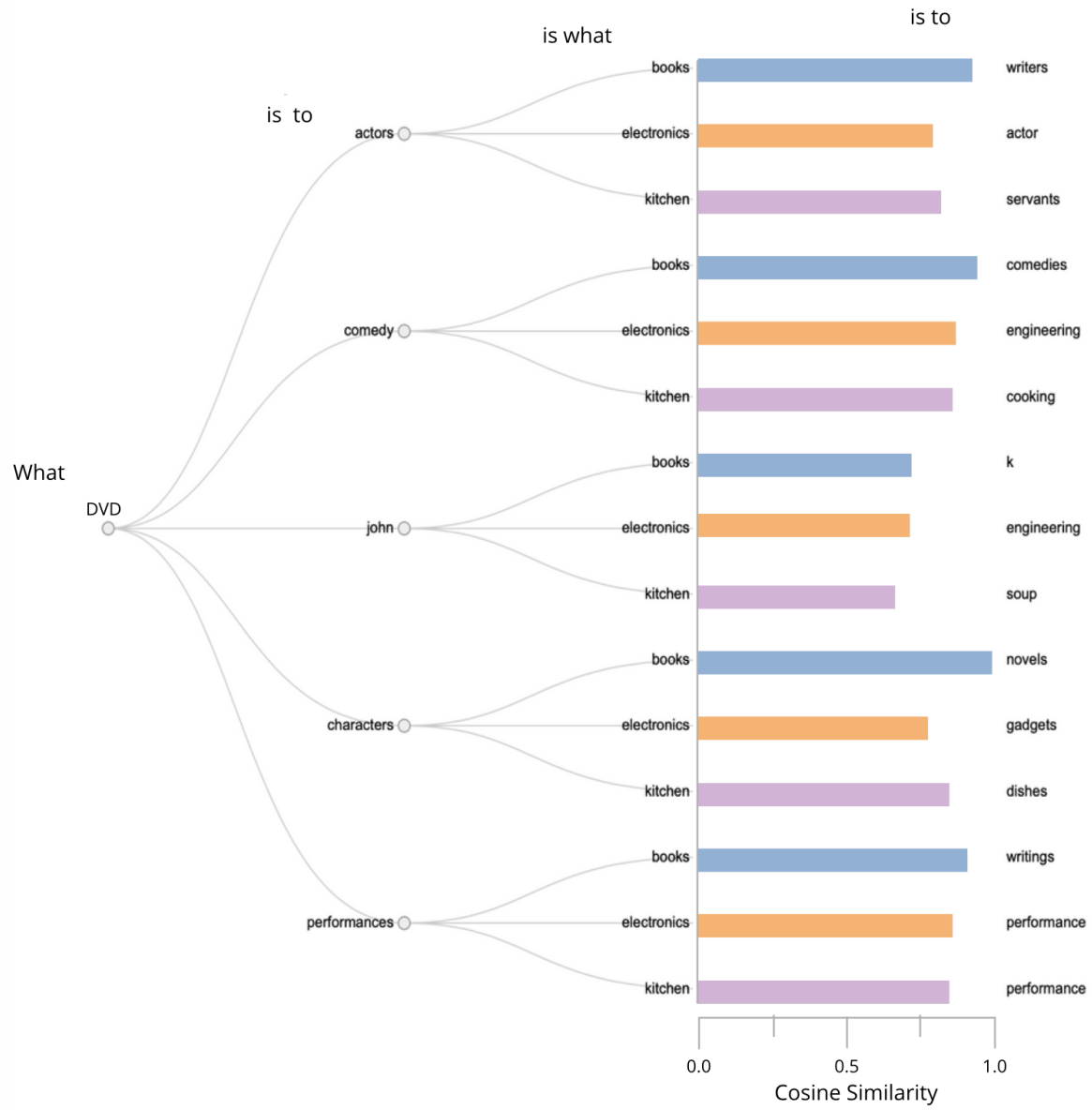


Figure 5: This figure visualizes the final mappings from the top 5 scoring words in "dvd", to the three target domains "books", "electronics" and "kitchen". Furthermore, this visualization also shows the cosine simularity which can be interpreted as how confident we are in the the mapping.

## References


[1] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. In *COLING*, 2018.

[2] Alberto Cairo. *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders Publishing, Thousand Oaks, CA, USA, 1st edition, 2016. ISBN 0321934075, 9780321934079.

[3] T Mikolov, W.-T Yih, and G Zweig. Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT*, pages 746–751, 01 2013.

[4] Donald A. Norman. *The Design of Everyday Things*. Basic Books, Inc., New York, NY, USA, 2002. ISBN 9780465067107.

[5] Juan Ramos. Using tf-idf to determine word relevance in document queries. 01 2003.