

PROJECT #1

Text representation with Higher-Order Paths

Advantages of using higher-order paths [1] between documents are illustrated in Fig. 1. In this figure, there are three documents, d_1 , d_2 and d_3 which include a set of terms $\{t_1, t_2, t_3\}$, $\{t_3, t_4, t_5\}$ and $\{t_4, t_5\}$ respectively. Using a traditional similarity measure which is based on the shared terms (e.g. dot product), similarity value between documents d_1 and d_3 will be zero since they do not share any terms. But in fact these two documents have some similarities in the context of the dataset through d_2 as it can be seen in Fig. 1. This supports the idea that using higher-order paths between documents, it is possible to obtain a non-zero similarity value between d_1 and d_3 which was not possible in traditional Bag of Words (BOW) [2] representation. This value becomes larger if there are many interconnecting documents like d_2 between d_1 and d_3 . This may stem from the reason that the two documents are written on the same topic using two different but semantically closer sets of terms.

This project aims to represent these higher-order paths by using *Linked-Lists*. Consequently, this project is a programming assignment in C, which aims to build an algorithm based on linked-lists that will build an efficient representation of documents.

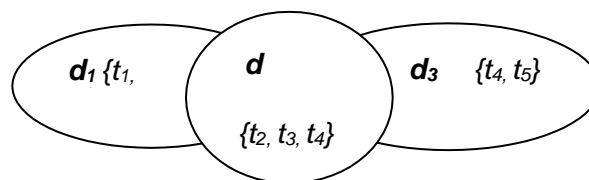
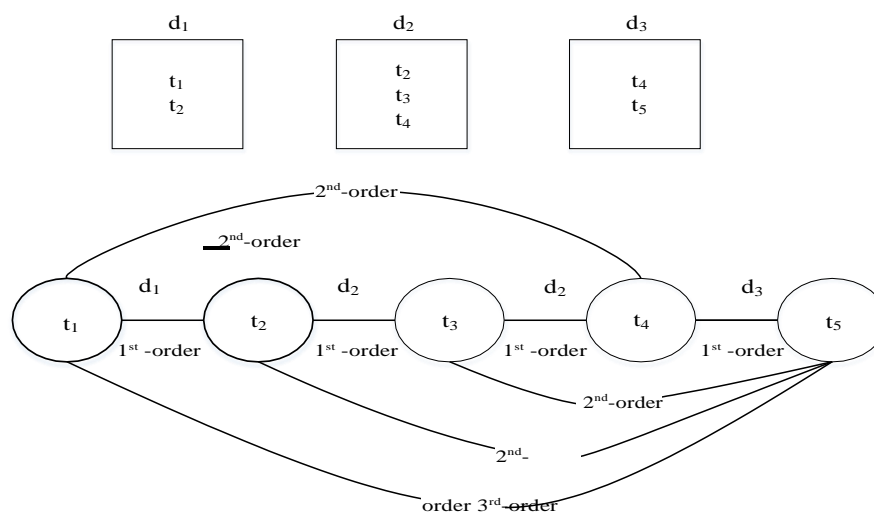


Fig. 1. a) Illustration of higher-order paths



1st-order term co-occurrence $\{t_1, t_2\}, \{t_2, t_3\}, \{t_3, t_4\}, \{t_2, t_4\}, \{t_4, t_5\}$

2nd-order term co-occurrence $\{t_1, t_3\}, \{t_1, t_4\}, \{t_2, t_5\}, \{t_3, t_5\}$

3rd-order term co-occurrence $\{t_1, t_5\}$

Fig. 2. b) Graphical demonstration of first-order, second-order and third-order paths between terms through documents

Your program needs to open and read text files under the following directories: econ, magazine and health. These are 3 categories of 1150Haber dataset [3]. The number of documents in these categories will be arbitrary. Furthermore, the number of terms in these documents will also be arbitrary. In other words, the length of these files will be arbitrary.

Your program is expected to do the followings:

- a) (60 points) You need to read all the documents under all the categories. Then you need to build a Linked-List (MLL). Each node in this MLL needs to represent a different term in these documents. All the terms in these documents are expected to be in the MLL. There will be cases, the same word occur in different documents, or in the same document. Then, you do not need to add a term into the MLL if it already exists. In other words, be careful about not entering the duplicate records into the MLL. After reading and storing all your data into Linked list, you are expected to find 1st, 2nd and 3rd order term co-occurrences as shown below.

1st-order term co-occurrence {t₁, t₂}, {t₂, t₃}, {t₃, t₄}, {t₄, t₅}, {t₅, t₆}, {t₆, t₇}, {t₇, t₈}, {t₈, t₉}, {t₉, t₁₀}, {t₁₀, t₁₁}, {t₁₁, t₁₂}, {t₁₂, t₁₃}, {t₁₃, t₁₄}, {t₁₄, t₁₅}, {t₁₅, t₁₆}, {t₁₆, t₁₇}, {t₁₇, t₁₈}, {t₁₈, t₁₉}, {t₁₉, t₂₀}, {t₂₀, t₂₁}, {t₂₁, t₂₂}, {t₂₂, t₂₃}, {t₂₃, t₂₄}, {t₂₄, t₂₅}, {t₂₅, t₂₆}, {t₂₆, t₂₇}, {t₂₇, t₂₈}, {t₂₈, t₂₉}, {t₂₉, t₃₀}, {t₃₀, t₃₁}, {t₃₁, t₃₂}, {t₃₂, t₃₃}, {t₃₃, t₃₄}, {t₃₄, t₃₅}, {t₃₅, t₃₆}, {t₃₆, t₃₇}, {t₃₇, t₃₈}, {t₃₈, t₃₉}, {t₃₉, t₄₀}, {t₄₀, t₄₁}, {t₄₁, t₄₂}, {t₄₂, t₄₃}, {t₄₃, t₄₄}, {t₄₄, t₄₅}, {t₄₅, t₄₆}, {t₄₆, t₄₇}, {t₄₇, t₄₈}, {t₄₈, t₄₉}, {t₄₉, t₅₀}, {t₅₀, t₅₁}, {t₅₁, t₅₂}, {t₅₂, t₅₃}, {t₅₃, t₅₄}, {t₅₄, t₅₅}, {t₅₅, t₅₆}, {t₅₆, t₅₇}, {t₅₇, t₅₈}, {t₅₈, t₅₉}, {t₅₉, t₆₀}, {t₆₀, t₆₁}, {t₆₁, t₆₂}, {t₆₂, t₆₃}, {t₆₃, t₆₄}, {t₆₄, t₆₅}, {t₆₅, t₆₆}, {t₆₆, t₆₇}, {t₆₇, t₆₈}, {t₆₈, t₆₉}, {t₆₉, t₇₀}, {t₇₀, t₇₁}, {t₇₁, t₇₂}, {t₇₂, t₇₃}, {t₇₃, t₇₄}, {t₇₄, t₇₅}, {t₇₅, t₇₆}, {t₇₆, t₇₇}, {t₇₇, t₇₈}, {t₇₈, t₇₉}, {t₇₉, t₈₀}, {t₈₀, t₈₁}, {t₈₁, t₈₂}, {t₈₂, t₈₃}, {t₈₃, t₈₄}, {t₈₄, t₈₅}, {t₈₅, t₈₆}, {t₈₆, t₈₇}, {t₈₇, t₈₈}, {t₈₈, t₈₉}, {t₈₉, t₉₀}, {t₉₀, t₉₁}, {t₉₁, t₉₂}, {t₉₂, t₉₃}, {t₉₃, t₉₄}, {t₉₄, t₉₅}, {t₉₅, t₉₆}, {t₉₆, t₉₇}, {t₉₇, t₉₈}, {t₉₈, t₉₉}, {t₉₉, t₁₀₀}, {t₁₀₀, t₁₀₁}, {t₁₀₁, t₁₀₂}, {t₁₀₂, t₁₀₃}, {t₁₀₃, t₁₀₄}, {t₁₀₄, t₁₀₅}, {t₁₀₅, t₁₀₆}, {t₁₀₆, t₁₀₇}, {t₁₀₇, t₁₀₈}, {t₁₀₈, t₁₀₉}, {t₁₀₉, t₁₁₀}, {t₁₁₀, t₁₁₁}, {t₁₁₁, t₁₁₂}, {t₁₁₂, t₁₁₃}, {t₁₁₃, t₁₁₄}, {t₁₁₄, t₁₁₅}, {t₁₁₅, t₁₁₆}, {t₁₁₆, t₁₁₇}, {t₁₁₇, t₁₁₈}, {t₁₁₈, t₁₁₉}, {t₁₁₉, t₁₂₀}, {t₁₂₀, t₁₂₁}, {t₁₂₁, t₁₂₂}, {t₁₂₂, t₁₂₃}, {t₁₂₃, t₁₂₄}, {t₁₂₄, t₁₂₅}, {t₁₂₅, t₁₂₆}, {t₁₂₆, t₁₂₇}, {t₁₂₇, t₁₂₈}, {t₁₂₈, t₁₂₉}, {t₁₂₉, t₁₃₀}, {t₁₃₀, t₁₃₁}, {t₁₃₁, t₁₃₂}, {t₁₃₂, t₁₃₃}, {t₁₃₃, t₁₃₄}, {t₁₃₄, t₁₃₅}, {t₁₃₅, t₁₃₆}, {t₁₃₆, t₁₃₇}, {t₁₃₇, t₁₃₈}, {t₁₃₈, t₁₃₉}, {t₁₃₉, t₁₄₀}, {t₁₄₀, t₁₄₁}, {t₁₄₁, t₁₄₂}, {t₁₄₂, t₁₄₃}, {t₁₄₃, t₁₄₄}, {t₁₄₄, t₁₄₅}, {t₁₄₅, t₁₄₆}, {t₁₄₆, t₁₄₇}, {t₁₄₇, t₁₄₈}, {t₁₄₈, t₁₄₉}, {t₁₄₉, t₁₅₀}, {t₁₅₀, t₁₅₁}, {t₁₅₁, t₁₅₂}, {t₁₅₂, t₁₅₃}, {t₁₅₃, t₁₅₄}, {t₁₅₄, t₁₅₅}, {t₁₅₅, t₁₅₆}, {t₁₅₆, t₁₅₇}, {t₁₅₇, t₁₅₈}, {t₁₅₈, t₁₅₉}, {t₁₅₉, t₁₆₀}, {t₁₆₀, t₁₆₁}, {t₁₆₁, t₁₆₂}, {t₁₆₂, t₁₆₃}, {t₁₆₃, t₁₆₄}, {t₁₆₄, t₁₆₅}, {t₁₆₅, t₁₆₆}, {t₁₆₆, t₁₆₇}, {t₁₆₇, t₁₆₈}, {t₁₆₈, t₁₆₉}, {t₁₆₉, t₁₇₀}, {t₁₇₀, t₁₇₁}, {t₁₇₁, t₁₇₂}, {t₁₇₂, t₁₇₃}, {t₁₇₃, t₁₇₄}, {t₁₇₄, t₁₇₅}, {t₁₇₅, t₁₇₆}, {t₁₇₆, t₁₇₇}, {t₁₇₇, t₁₇₈}, {t₁₇₈, t₁₇₉}, {t₁₇₉, t₁₈₀}, {t₁₈₀, t₁₈₁}, {t₁₈₁, t₁₈₂}, {t₁₈₂, t₁₈₃}, {t₁₈₃, t₁₈₄}, {t₁₈₄, t₁₈₅}, {t₁₈₅, t₁₈₆}, {t₁₈₆, t₁₈₇}, {t₁₈₇, t₁₈₈}, {t₁₈₈, t₁₈₉}, {t₁₈₉, t₁₉₀}, {t₁₉₀, t₁₉₁}, {t₁₉₁, t₁₉₂}, {t₁₉₂, t₁₉₃}, {t₁₉₃, t₁₉₄}, {t₁₉₄, t₁₉₅}, {t₁₉₅, t₁₉₆}, {t₁₉₆, t₁₉₇}, {t₁₉₇, t₁₉₈}, {t₁₉₈, t₁₉₉}, {t₁₉₉, t₂₀₀}, {t₂₀₀, t₂₀₁}, {t₂₀₁, t₂₀₂}, {t₂₀₂, t₂₀₃}, {t₂₀₃, t₂₀₄}, {t₂₀₄, t₂₀₅}, {t₂₀₅, t₂₀₆}, {t₂₀₆, t₂₀₇}, {t₂₀₇, t₂₀₈}, {t₂₀₈, t₂₀₉}, {t₂₀₉, t₂₁₀}, {t₂₁₀, t₂₁₁}, {t₂₁₁, t₂₁₂}, {t₂₁₂, t₂₁₃}, {t₂₁₃, t₂₁₄}, {t₂₁₄, t₂₁₅}, {t₂₁₅, t₂₁₆}, {t₂₁₆, t₂₁₇}, {t₂₁₇, t₂₁₈}, {t₂₁₈, t₂₁₉}, {t₂₁₉, t₂₂₀}, {t₂₂₀, t₂₂₁}, {t₂₂₁, t₂₂₂}, {t₂₂₂, t₂₂₃}, {t₂₂₃, t₂₂₄}, {t₂₂₄, t₂₂₅}, {t₂₂₅, t₂₂₆}, {t₂₂₆, t₂₂₇}, {t₂₂₇, t₂₂₈}, {t₂₂₈, t₂₂₉}, {t₂₂₉, t₂₃₀}, {t₂₃₀, t₂₃₁}, {t₂₃₁, t₂₃₂}, {t₂₃₂, t₂₃₃}, {t₂₃₃, t₂₃₄}, {t₂₃₄, t₂₃₅}, {t₂₃₅, t₂₃₆}, {t₂₃₆, t₂₃₇}, {t₂₃₇, t₂₃₈}, {t₂₃₈, t₂₃₉}, {t₂₃₉, t₂₄₀}, {t₂₄₀, t₂₄₁}, {t₂₄₁, t₂₄₂}, {t₂₄₂, t₂₄₃}, {t₂₄₃, t₂₄₄}, {t₂₄₄, t₂₄₅}, {t₂₄₅, t₂₄₆}, {t₂₄₆, t₂₄₇}, {t₂₄₇, t₂₄₈}, {t₂₄₈, t₂₄₉}, {t₂₄₉, t₂₅₀}, {t₂₅₀, t₂₅₁}, {t₂₅₁, t₂₅₂}, {t₂₅₂, t₂₅₃}, {t₂₅₃, t₂₅₄}, {t₂₅₄, t₂₅₅}, {t₂₅₅, t₂₅₆}, {t₂₅₆, t₂₅₇}, {t₂₅₇, t₂₅₈}, {t₂₅₈, t₂₅₉}, {t₂₅₉, t₂₆₀}, {t₂₆₀, t₂₆₁}, {t₂₆₁, t₂₆₂}, {t₂₆₂, t₂₆₃}, {t₂₆₃, t₂₆₄}, {t₂₆₄, t₂₆₅}, {t₂₆₅, t₂₆₆}, {t₂₆₆, t₂₆₇}, {t₂₆₇, t₂₆₈}, {t₂₆₈, t₂₆₉}, {t₂₆₉, t₂₇₀}, {t₂₇₀, t₂₇₁}, {t₂₇₁, t₂₇₂}, {t₂₇₂, t₂₇₃}, {t₂₇₃, t₂₇₄}, {t₂₇₄, t₂₇₅}, {t₂₇₅, t₂₇₆}, {t₂₇₆, t₂₇₇}, {t₂₇₇, t₂₇₈}, {t₂₇₈, t₂₇₉}, {t₂₇₉, t₂₈₀}, {t₂₈₀, t₂₈₁}, {t₂₈₁, t₂₈₂}, {t₂₈₂, t₂₈₃}, {t₂₈₃, t₂₈₄}, {t₂₈₄, t₂₈₅}, {t₂₈₅, t₂₈₆}, {t₂₈₆, t₂₈₇}, {t₂₈₇, t₂₈₈}, {t₂₈₈, t₂₈₉}, {t₂₈₉, t₂₉₀}, {t₂₉₀, t₂₉₁}, {t₂₉₁, t₂₉₂}, {t₂₉₂, t₂₉₃}, {t₂₉₃, t₂₉₄}, {t₂₉₄, t₂₉₅}, {t₂₉₅, t₂₉₆}, {t₂₉₆, t₂₉₇}, {t₂₉₇, t₂₉₈}, {t₂₉₈, t₂₉₉}, {t₂₉₉, t₃₀₀}, {t₃₀₀, t₃₀₁}, {t₃₀₁, t₃₀₂}, {t₃₀₂, t₃₀₃}, {t₃₀₃, t₃₀₄}, {t₃₀₄, t₃₀₅}, {t₃₀₅, t₃₀₆}, {t₃₀₆, t₃₀₇}, {t₃₀₇, t₃₀₈}, {t₃₀₈, t₃₀₉}, {t₃₀₉, t₃₁₀}, {t₃₁₀, t₃₁₁}, {t₃₁₁, t₃₁₂}, {t₃₁₂, t₃₁₃}, {t₃₁₃, t₃₁₄}, {t₃₁₄, t₃₁₅}, {t₃₁₅, t₃₁₆}, {t₃₁₆, t₃₁₇}, {t₃₁₇, t₃₁₈}, {t₃₁₈, t₃₁₉}, {t₃₁₉, t₃₂₀}, {t₃₂₀, t₃₂₁}, {t₃₂₁, t₃₂₂}, {t₃₂₂, t₃₂₃}, {t₃₂₃, t₃₂₄}, {t₃₂₄, t₃₂₅}, {t₃₂₅, t₃₂₆}, {t₃₂₆, t₃₂₇}, {t₃₂₇, t₃₂₈}, {t₃₂₈, t₃₂₉}, {t₃₂₉, t₃₃₀}, {t₃₃₀, t₃₃₁}, {t₃₃₁, t₃₃₂}, {t₃₃₂, t₃₃₃}, {t₃₃₃, t₃₃₄}, {t₃₃₄, t₃₃₅}, {t₃₃₅, t₃₃₆}, {t₃₃₆, t₃₃₇}, {t₃₃₇, t₃₃₈}, {t₃₃₈, t₃₃₉}, {t₃₃₉, t₃₄₀}, {t₃₄₀, t₃₄₁}, {t₃₄₁, t₃₄₂}, {t₃₄₂, t₃₄₃}, {t₃₄₃, t₃₄₄}, {t₃₄₄, t₃₄₅}, {t₃₄₅, t₃₄₆}, {t₃₄₆, t₃₄₇}, {t₃₄₇, t₃₄₈}, {t₃₄₈, t₃₄₉}, {t₃₄₉, t₃₅₀}, {t₃₅₀, t₃₅₁}, {t₃₅₁, t₃₅₂}, {t₃₅₂, t₃₅₃}, {t₃₅₃, t₃₅₄}, {t₃₅₄, t₃₅₅}, {t₃₅₅, t₃₅₆}, {t₃₅₆, t₃₅₇}, {t₃₅₇, t₃₅₈}, {t₃₅₈, t₃₅₉}, {t₃₅₉, t₃₆₀}, {t₃₆₀, t₃₆₁}, {t₃₆₁, t₃₆₂}, {t₃₆₂, t₃₆₃}, {t₃₆₃, t₃₆₄}, {t₃₆₄, t₃₆₅}, {t₃₆₅, t₃₆₆}, {t₃₆₆, t₃₆₇}, {t₃₆₇, t₃₆₈}, {t₃₆₈, t₃₆₉}, {t₃₆₉, t₃₇₀}, {t₃₇₀, t₃₇₁}, {t₃₇₁, t₃₇₂}, {t₃₇₂, t₃₇₃}, {t₃₇₃, t₃₇₄}, {t₃₇₄, t₃₇₅}, {t₃₇₅, t₃₇₆}, {t₃₇₆, t₃₇₇}, {t₃₇₇, t₃₇₈}, {t₃₇₈, t₃₇₉}, {t₃₇₉, t₃₈₀}, {t₃₈₀, t₃₈₁}, {t₃₈₁, t₃₈₂}, {t₃₈₂, t₃₈₃}, {t₃₈₃, t₃₈₄}, {t₃₈₄, t₃₈₅}, {t₃₈₅, t₃₈₆}, {t₃₈₆, t₃₈₇}, {t₃₈₇, t₃₈₈}, {t₃₈₈, t₃₈₉}, {t₃₈₉, t₃₉₀}, {t₃₉₀, t₃₉₁}, {t₃₉₁, t₃₉₂}, {t₃₉₂, t₃₉₃}, {t₃₉₃, t₃₉₄}, {t₃₉₄, t₃₉₅}, {t₃₉₅, t₃₉₆}, {t₃₉₆, t₃₉₇}, {t₃₉₇, t₃₉₈}, {t₃₉₈, t₃₉₉}, {t₃₉₉, t₄₀₀}, {t₄₀₀, t₄₀₁}, {t₄₀₁, t₄₀₂}, {t₄₀₂, t₄₀₃}, {t₄₀₃, t₄₀₄}, {t₄₀₄, t₄₀₅}, {t₄₀₅, t₄₀₆}, {t₄₀₆, t₄₀₇}, {t₄₀₇, t₄₀₈}, {t₄₀₈, t₄₀₉}, {t₄₀₉, t₄₁₀}, {t₄₁₀, t₄₁₁}, {t₄₁₁, t₄₁₂}, {t₄₁₂, t₄₁₃}, {t₄₁₃, t₄₁₄}, {t₄₁₄, t₄₁₅}, {t₄₁₅, t₄₁₆}, {t₄₁₆, t₄₁₇}, {t₄₁₇, t₄₁₈}, {t₄₁₈, t₄₁₉}, {t₄₁₉, t₄₂₀}, {t₄₂₀, t₄₂₁}, {t₄₂₁, t₄₂₂}, {t₄₂₂, t₄₂₃}, {t₄₂₃, t₄₂₄}, {t₄₂₄, t₄₂₅}, {t₄₂₅, t₄₂₆}, {t₄₂₆, t₄₂₇}, {t₄₂₇, t₄₂₈}, {t₄₂₈, t₄₂₉}, {t₄₂₉, t₄₃₀}, {t₄₃₀, t₄₃₁}, {t₄₃₁, t₄₃₂}, {t₄₃₂, t₄₃₃}, {t₄₃₃, t₄₃₄}, {t₄₃₄, t₄₃₅}, {t₄₃₅, t₄₃₆}, {t₄₃₆, t₄₃₇}, {t₄₃₇, t₄₃₈}, {t₄₃₈, t₄₃₉}, {t₄₃₉, t₄₄₀}, {t₄₄₀, t₄₄₁}, {t₄₄₁, t₄₄₂}, {t₄₄₂, t₄₄₃}, {t₄₄₃, t₄₄₄}, {t₄₄₄, t₄₄₅}, {t₄₄₅, t₄₄₆}, {t₄₄₆, t₄₄₇}, {t₄₄₇, t₄₄₈}, {t₄₄₈, t₄₄₉}, {t₄₄₉, t₄₅₀}, {t₄₅₀, t₄₅₁}, {t₄₅₁, t₄₅₂}, {t₄₅₂, t₄₅₃}, {t₄₅₃, t₄₅₄}, {t₄₅₄, t₄₅₅}, {t₄₅₅, t₄₅₆}, {t₄₅₆, t₄₅₇}, {t₄₅₇, t₄₅₈}, {t₄₅₈, t₄₅₉}, {t₄₅₉, t₄₆₀}, {t₄₆₀, t₄₆₁}, {t₄₆₁, t₄₆₂}, {t₄₆₂, t₄₆₃}, {t₄₆₃, t₄₆₄}, {t₄₆₄, t₄₆₅}, {t₄₆₅, t₄₆₆}, {t₄₆₆, t₄₆₇}, {t₄₆₇, t₄₆₈}, {t₄₆₈, t₄₆₉}, {t₄₆₉, t₄₇₀}, {t₄₇₀, t₄₇₁}, {t₄₇₁, t₄₇₂}, {t₄₇₂, t₄₇₃}, {t₄₇₃, t₄₇₄}, {t₄₇₄, t₄₇₅}, {t₄₇₅, t₄₇₆}, {t₄₇₆, t₄₇₇}, {t₄₇₇, t₄₇₈}, {t₄₇₈, t₄₇₉}, {t₄₇₉, t₄₈₀}, {t₄₈₀, t₄₈₁}, {t₄₈₁, t₄₈₂}, {t₄₈₂, t₄₈₃}, {t₄₈₃, t₄₈₄}, {t₄₈₄, t₄₈₅}, {t₄₈₅, t₄₈₆}, {t₄₈₆, t₄₈₇}, {t₄₈₇, t₄₈₈}, {t₄₈₈, t₄₈₉}, {t₄₈₉, t₄₉₀}, {t₄₉₀, t₄₉₁}, {t₄₉₁, t₄₉₂}, {t₄₉₂, t₄₉₃}, {t₄₉₃, t₄₉₄}, {t₄₉₄, t₄₉₅}, {t₄₉₅, t₄₉₆}, {t₄₉₆, t₄₉₇}, {t₄₉₇, t₄₉₈}, {t₄₉₈, t₄₉₉}, {t₄₉₉, t₅₀₀}, {t₅₀₀, t₅₀₁}, {t₅₀₁, t₅₀₂}, {t₅₀₂, t₅₀₃}, {t₅₀₃, t₅₀₄}, {t₅₀₄, t₅₀₅}, {t₅₀₅, t₅₀₆}, {t₅₀₆, t₅₀₇}, {t₅₀₇, t₅₀₈}, {t₅₀₈, t₅₀₉}, {t₅₀₉, t₅₁₀}, {t₅₁₀, t₅₁₁}, {t₅₁₁, t₅₁₂}, {t₅₁₂, t₅₁₃}, {t₅₁₃, t₅₁₄}, {t₅₁₄, t₅₁₅}, {t₅₁₅, t₅₁₆}, {t₅₁₆, t₅₁₇}, {t₅₁₇, t₅₁₈}, {t₅₁₈, t₅₁₉}, {t₅₁₉, t₅₂₀}, {t₅₂₀, t₅₂₁}, {t₅₂₁, t₅₂₂}, {t₅₂₂, t₅₂₃}, {t₅₂₃, t₅₂₄}, {t₅₂₄, t₅₂₅}, {t₅₂₅, t₅₂₆}, {t₅₂₆, t₅₂₇}, {t₅₂₇, t₅₂₈}, {t₅₂₈, t₅₂₉}, {t₅₂₉, t₅₃₀}, {t₅₃₀, t₅₃₁}, {t₅₃₁, t₅₃₂}, {t₅₃₂, t₅₃₃}, {t₅₃₃, t₅₃₄}, {t₅₃₄, t₅₃₅}, {t₅₃₅, t₅₃₆}, {t₅₃₆, t₅₃₇}, {t₅₃₇, t₅₃₈}, {t₅₃₈, t₅₃₉}, {t₅₃₉, t₅₄₀}, {t₅₄₀, t₅₄₁}, {t₅₄₁, t₅₄₂}, {t₅₄₂, t₅₄₃}, {t₅₄₃, t₅₄₄}, {t₅₄₄, t₅₄₅}, {t₅₄₅, t₅₄₆}, {t₅₄₆, t₅₄₇}, {t₅₄₇, t₅₄₈}, {t₅₄₈, t₅₄₉}, {t₅₄₉, t₅₅₀}, {t₅₅₀, t₅₅₁}, {t₅₅₁, t₅₅₂}, {t₅₅₂, t₅₅₃}, {t₅₅₃, t₅₅₄}, {t₅₅₄, t₅₅₅}, {t

c) (20 points)

Output: Most frequent 10 words in the input set of documents *for each category*, sorted descending by their term frequency*inverse document frequency (idf) coupled with their tf-idf values.

$$\text{IDF}(t) = \log_e N/n$$

N : Total number of documents

n : Number of documents with term t in it.

Econ	Health	Magazine
Dolar,1.8	Operation,2.12	Cinema,3.24
Bank,1.7	Medicine,2.10	Actor,3.21
Strategy,0.6	Doctor,1.8	Theatre,2.18
...

REFERENCES

- [1] Altinel, B., Ganiz, M. C., & Diri, B., 2014 . A semantic kernel for text classification based on iterative higher-order relations between words and documents. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 505-517). Springer, Cham.
- [2] Salton, G., Yang, C.S., 1973. On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, 29(4):11-21.
- [3] Amasyalı, M. F. and Beken, A. Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması, in *Proc IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU)*, IEEE Press, 2009.

