International Hellenic University

Natural Language Processing and Text Mining
Academic Year 2022-2023
Assignment: “Relevance Prediction”

A. Introduction. The area of text analytics (or text mining) includes techniques from multitude scientific areas (A.I., statistics, linguistics), and it has a wide range of applications (security, marketing, information retrieval, opinion mining). While structured data are “ready” for analysis and evaluation, unstructured text requires transformation in order to uncover the underlying information. The transformation of unstructured text into a structured set of data is not a straight forward task and text analytics offers a wide variety of tools to tackle with the idioms, ambiguities and irregularities of natural language.

In this assignment, you are going to work in teams (2 persons per team). The task that must be performed is to predict the relevance of a result with respect to a query. Each query is represented by a search term or terms and the result is a specific product. Based on the relevance between the query and the answer, a relevance score is assigned. The higher the score the better the relevance of the answer. The relevance is a real number in the range [1,3] (in steps of 0.5), where 1 denotes minimum relevance and 3 is the maximum relevance. Your job is to be able to predict the relevance score for an unknown combination between a query and a result. The evaluation measure used is the **Root-Mean-Square Error (RMSE)**.

B. Datasets. There are three datasets available. The first one, **train.csv**, contains pairs of queries and answers and also contains the relevance score. A screenshot of the dataset is shown in Figure 1. As you can observe, a product may be the best answer for different queries (search terms).

id	product_uid	product_title	search_term	relevance
2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3
3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.5
9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 Tugboat Wood and Concrete Coating	deck over	3

Figure 1. Screenshot of **train.csv**.

The second dataset is **product_descriptions.csv**, and for each different product it contains a textual description of the product. A screenshot is shown in Figure 2.

```
"product uid","product description"
100001,"Not only do angles make joints stronger, they also provide more consistent, straight corners. Simpson Strong-Tie offers a wide variety of angles in various sizes and thicknesses to handle light-duty jobs or projects where a structural connection is needed. Some can be bent (skewed) to match the project. For outdoor projects or those where moisture is present, use our ZMAX zinc-coated connectors, which provide extra resistance against corrosion (look for a ""Z"" at the end of the model number).Versatile connector for various 90 connections and home repair projectsStronger than angled nailing or screw fastening aloneHelp ensure joints are consistently straight and strongDimensions: 3 in. x 3 in. x 1-1/2 in.Made from 12-Gauge steelGalvanized for extra corrosion resistanceInstall with 10d common nails or #9 x 1-1/2 in. Strong-Drive SD screws"
100002,"BEHR Premium Textured DECKOVER is an innovative solid color coating. It will bring your old, weathered wood or concrete back to life. The advanced 100% acrylic resin formula creates a durable coating for your tired and worn out deck, rejuvenating to a whole new look.  For the best results, be sure to properly prepare the surface using other applicable BEHR products displayed above.California residents: see Proposition 65 informationRevives wood and composite decks, railings, porches and boat docks, also great for concrete pool decks, patios and sidewalks100% acrylic solid color coatingResists cracking and peeling and conceals splinters and cracks up to 1/4 in.Provides a durable, mildew resistant finishCovers up to 75 sq. ft. in 2 coats per gallonCreates a textured, slip-resistant finishFor best results, prepare with the appropriate BEHR product for your wood or concrete surfaceActual paint colors may vary from on-screen and printer representationsColors available to be tinted in most storesOnline Price includes Paint Care fee in the following states: CA, CO, CT, ME, MN, OR, RI, VT"
```

Figure 2. Screenshot of **product_descriptions.csv**.

```
"product uid","name","value"
100001,"Bullet01","Versatile connector for various 90° connections and home repair projects"
100001,"Bullet02","Stronger than angled nailing or screw fastening alone"
100001,"Bullet03","Help ensure joints are consistently straight and strong"
100001,"Bullet04","Dimensions: 3 in. x 3 in. x 1-1/2 in."
100001,"Bullet05","Made from 12-Gauge steel"
100001,"Bullet06","Galvanized for extra corrosion resistance"
100001,"Bullet07","Install with 10d common nails or #9 x 1-1/2 in. Strong-Drive SD screws"
100001,"Gauge","12"
100001,"Material","Galvanized Steel"
100001,"MFG Brand Name","Simpson Strong-Tie"
100001,"Number of Pieces","1"
100001,"Product Depth (in.)","1.5"
100001,"Product Height (in.)","3"
100001,"Product Weight (lb.)","0.26"
100001,"Product Width (in.)","3"
100002,"Application Method","Brush,Roller,Spray"
100002,"Assembled Depth (in.)","6.63 in"
100002,"Assembled Height (in.)","7.76 in"
100002,"Assembled Width (in.)","6.63 in"
```

Figure 3. Screenshot of **attributes.csv**.

The third file, **attributes.csv**, contains additional attributes for some of the products (some products may not contain additional attributes). A screenshot is shown in Figure 3.

C. Deliverables. You are requested to write code in any programming language (preferably in Python) in order to solve the relevance prediction problem. You are free to use supervised or unsupervised learning techniques, use feature extraction techniques for text, apply text preprocessing, and in general you are free to apply any technique aiming at the **minimization of RMSE**. In addition to the source code, you must prepare a technical report, no more than 5 pages, explaining your solution and the techniques you have tested, sets of features that have shown the best performance, feature engineering, etc. You should upload your solutions using the corresponding assignment link in the IHU elearning platform. The deadline for the submission is set to **June 18, 2023**. The assignment grade will contribute to **30%** of your total course grade.

Try to get the most out of this assignment, by applying techniques that you have learned or other methods that you may find in books and related material. Also, have fun with the project!

The NLP&TM Team