**PROGRAMMING ASSIGNMENT 1: WORD SIMILARITY**

**AND SEMANTIC RELATION CLASSIFICATION**

1. **Problem**
2. *Cosine similarity*: Given pre-trained embeddings of Vietnamese words, implement a function for calculating cosine similarity between word pairs. Test your program using word pairs in ViSim-400 dataset (in directory Datasets/ViSim-400). Using Pearson correlation coefficient (*https://en.wikipedia.org/wiki/Pearson\_correlation\_coefficient*), Spearman's rank correlation coefficient (*https://en.wikipedia.org/wiki/**Spearman%27s\_rank\_correlation\_coefficient*) to evaluate the correlation between your results and similarity scores in the dataset (See more example code in directory: *Word-Similarity*

*\Code Sample Python*).

1. *K-nearest words*: Given a word w, find k most-similar words of w using the function implemented in question 1.
2. *Synonym-antonym classification*: Implement a Logistic Regression classifier or a Multi-layer Perceptron classifier for distinguishing synonyms and antonyms. For training, you are given a data set in directory antonym-synonym set *(**https://github.com/NLP-Projects/Word-Similarity/tree/master/antonym-synonym%20set).* For testing, you use ViCon-400 (in directory Datasets/ViCon-400). Experimental results are evaluated by precision scores, recall scores and F-measure (F1).
3. **Programming language and tool**
4. Programming language: Python.
5. Machine learning library: Scikit-learn (https://scikit-learn.org/).
6. **Data sets**

Available at: ***<https://github.com/NLP-Projects/Word-Similarity>***

1. Pre-trained Word2Vec in directory *Word2vec*.

For example, the word “*sinh\_viên*” is represented by a 150-dimension vector as follows:

sinh\_viên -0.16830535 -0.46649584 -0.09095726 0.26220384 -0.06665505 0.031103149 -0.09404702 0.04053062 0.11192394 -0.5279207 0.14246537 0.001038614 -0.35036477 -0.13038214 -0.09090571 0.16573012 -0.36303198 0.09883489 0.18725859 0.13902836 0.06649695 -0.084366314 -0.0027223954 -0.19726162 0.2759428 0.19625933 -0.05050645 -0.2912832 0.3332136 -0.05097548 -0.033659954 0.08438698 0.08262109 0.17202266 0.028343566 -0.2239018 -0.12655334 -0.007960333 -0.11457155 0.05343645 0.04817521 -0.07958502 0.14042595 -0.03268163 0.15085939 -0.019562919 -0.3306483 -0.12484468 0.28407142 0.412389 -0.20573528 0.1886385 -0.32018003 -0.28494123 0.30086038 -0.16913833 -0.15604256 0.014038047 0.0027964886 -0.042381432 0.10412829 -0.25038096 -0.1707933 0.28154066 -0.18622164 -0.12197793 -0.3312451 -0.51899064 -0.30050278 -0.25337204 0.0281554 0.05594583 0.2137868 -0.063191235 -0.25344792 -0.041951556 -0.24166772 -0.06607595 -0.0058164373 0.20314898 0.22446826 -0.27940097 -0.20987804 0.393391 0.19083196 0.1140723 -0.0413766 0.2983006 0.09309805 -0.014998496 0.56122595 -0.21278302 -0.29513258 -0.5796372 -0.104330115 -0.049805988 -0.20145701 0.07974479 -0.08291912 0.20195685 0.5489658 0.27150062 0.28475645 0.047555167 -0.05718565 -0.25075287 0.16845582 -0.26866978 0.24444789 -0.1013144 0.34333876 0.44675854 0.2024813 -0.18543775 -0.428579 0.00044292794 -0.06001611 0.1679784 -0.18539493 0.5264743 0.032929808 -0.11656791 0.11542175 0.345688 -0.16840588 -0.10268665 0.16477033 -0.2467253 0.1251898 0.0076962546 -0.27712318 0.33789232 -0.1583204 0.19434617 -0.23928708 0.177563 0.11510917 0.2388904 -0.50243086 -0.53305346 0.01388552 0.18508705 0.28942367 0.10520081 -0.060460173 0.16218448 0.13140707 -0.16748051 -0.10331021 -0.17137058

1. VSim-400 dataset in directory *Datasets/ViSim-400*.

Line format and examples (Sim1 represents human rating of similarity in the interval [0,4], Sim2 represents human rating of similarity in the interval [0,6]):

Word1 Word2 POS Sim1 Sim2 STD

biến ngập V 3.13 5.22 0.72

động tĩnh V 0.6 1.0 0.95

khuyết ưu N 0.2 0.33 0.4

cõi\_tục cõi\_âm N 0.6 1.0 0.95

kết\_duyên thành\_hôn V 5.27 8.78 1.06

cấp\_tiến bảo\_thủ A 0.87 1.45 1.15

nước\_lớn nguy\_hiểm N 1.07 1.78 1.12

bất\_lợi thuận\_lợi N 0.33 0.55 0.79

phân\_ly sum\_họp V 0.47 0.78 1.09

diễu\_hành tuần\_hành V 4.53 7.55 1.15

1. Vcon-400 dataset in directory *Datasets/ViCon-400*.

Line format and examples:

Word1 Word2 Relation

hời\_hợt nông\_cạn SYN

thảnh\_thơi ưu\_tư ANT

vô\_lý có\_lí ANT

cuồng\_nộ phẫn\_nộ SYN

đần thông\_minh ANT

có\_lí vô\_lí ANT

ấm\_no no\_ấm SYN

lí\_thú lý\_thú SYN

1. Reference documents in directory *Reference*.