

TABLE OF CONTENETS

Dedication	II
ACKNOWLEDGEMENTS	III
Table of contenets	IV
List of figures	VII
List of Equations	IX
Chapter 1. General Introduction	2
1.1 Context:	2
1.2 Problematic:	3
1.3 Objectives:	3
1.4 Thesis plan:	3
Chapter 2. State of Art	6
2.1 Introduction:	6
2.2 Text similarity methods:	6
2.2.1 Corpus-based Similarity:	6
2.2.2 Knowledge-Based Similarity:	7
2.2.3 String-based Similarity:	8
2.2.3.1 Character-based Similarity:	9
2.2.3.2 <i>The term-based similarity:</i>	9

2.2.4	Hybrid Similarities:	10
2.3	Token-based models' objectives:	10
2.4	Conclusion:	11
Chapter 3. Architecture and Conception		13
3.1	Introduction:	13
3.2	Vector Space Model:	13
3.2.1	Definitions:	13
3.2.2	Models and approaches:	13
3.2.2.1	Inner Product:	14
3.2.2.2	Cosine Similarity:	15
3.2.2.3	Jaccard Similarity:	15
3.2.2.4	Dice Similarity:	16
3.2.3	Term Weighting:	16
3.2.3.1	TF-IDF weighting:	17
3.2.3.2	Term Frequency (TF):	17
3.2.3.3	Document frequency (DF):	17
3.2.3.4	Inverse Document Frequency (IDF):	18
3.2.3.5	Term frequency–Inverse document frequency(tf.idf):	18
3.3	Conception:	18
3.3.1	Tokenization:	19
3.3.2	Punctuation Removal:	19
3.3.3	Stop-word removal:	20
3.3.4	Lemmatization:	21
3.3.5	Creating words-bag:	22

3.3.6	Compute the term frequency:	22
3.4	Applying the vector space model Techniques:	23
3.5	Conclusion:	25
Chapter 4. Implementation		27
4.1	Introduction:	27
4.2	Programming Language:	27
4.2.1	Python Presentation:	27
4.2.2	Interactive Python:	28
4.3	Environment and tools:	28
4.3.1	Presentation of PyCharm:	28
4.3.2	PyQt presentation:	29
4.3.3	PyQt designer:	30
4.4	Application Presentation:	31
4.4.1	Main window:	32
4.4.2	English window:	32
4.4.3	French window:	36
4.5	Result discussion:	38
4.6	Conclusion:	39
General Conclusion		40
References		41