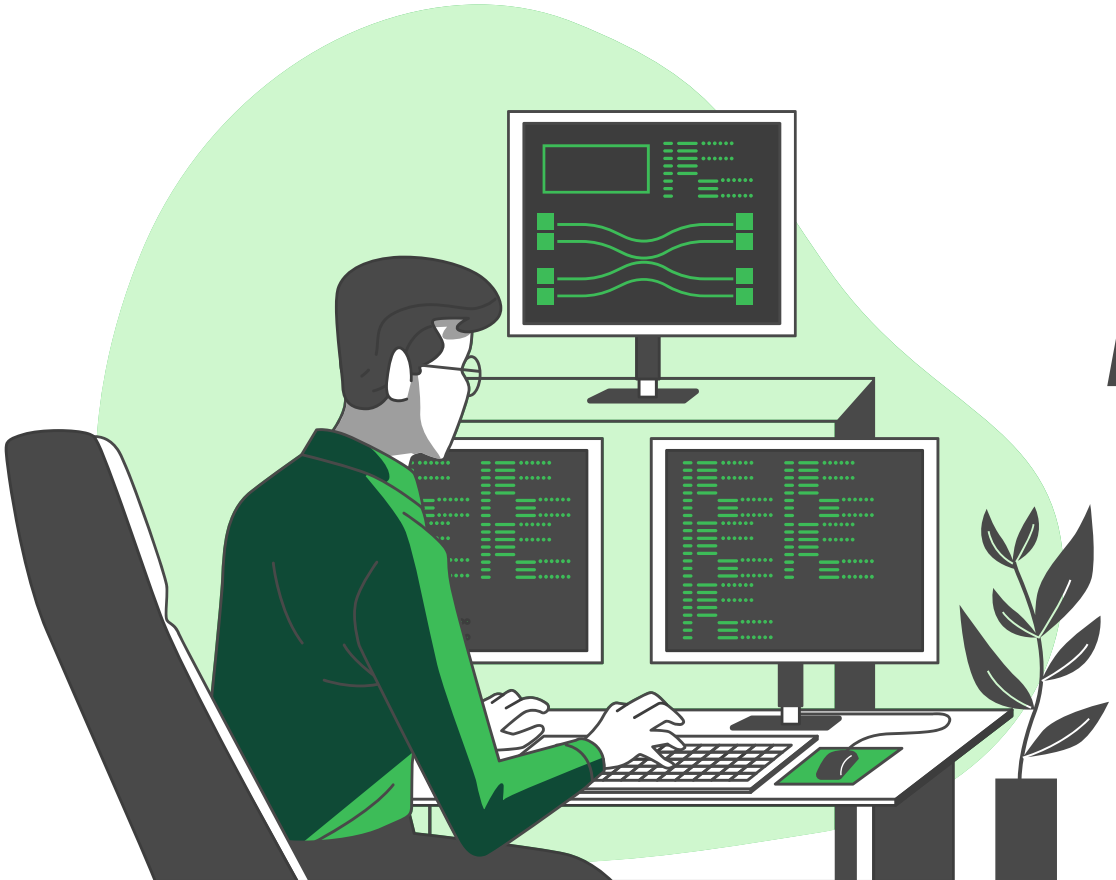


Similarity Analysis via NLP Models

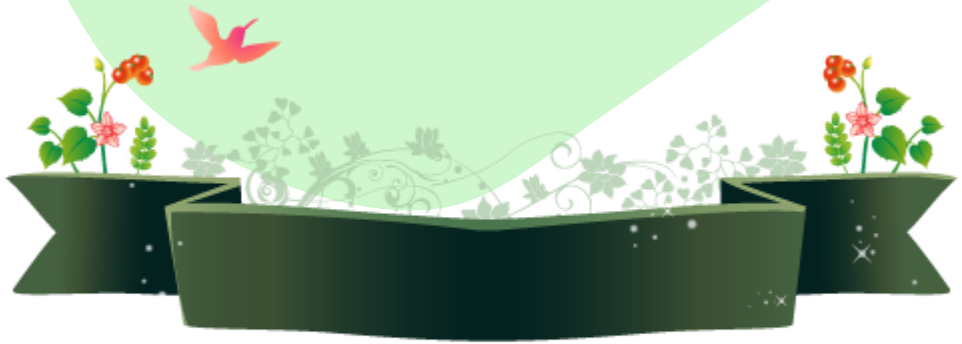
On Ahadis

Process

1. Preprocessing of available data
2. Vectorization of data
3. Similarity Computation
4. Hadith Equivalence class Generation
5. Result Analysis

01

Preprocessing



Understanding the Problem

Removing articles

Neptune is the farthest planet from the Sun

Removing Araab

أنا أحب أكل التفاح

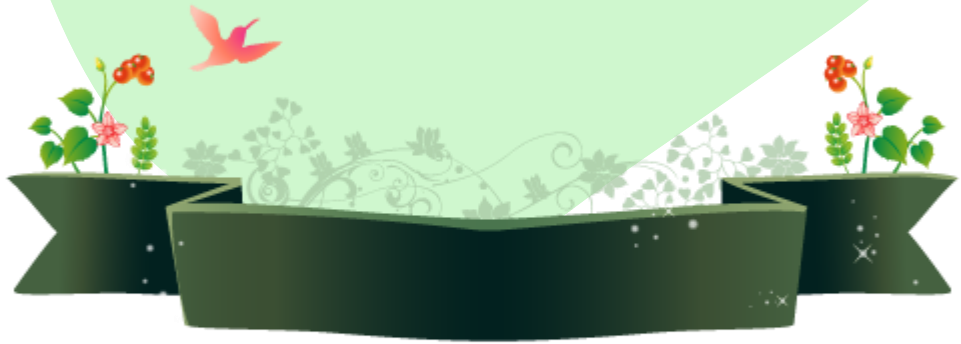
...

Checking for Data Inconsistency

...

02

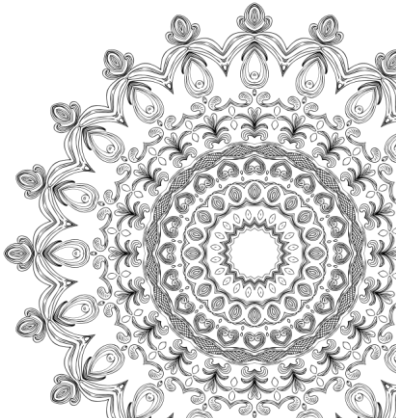
Vectorization



NLP Models

01 TDFIDF Model

02 BERT Model



TFIDF Model

A word frequency based model

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency

Count Vectorizer

	blue	bright	sky	sun
Doc1	1	0	1	0
Doc2	0	1	0	1

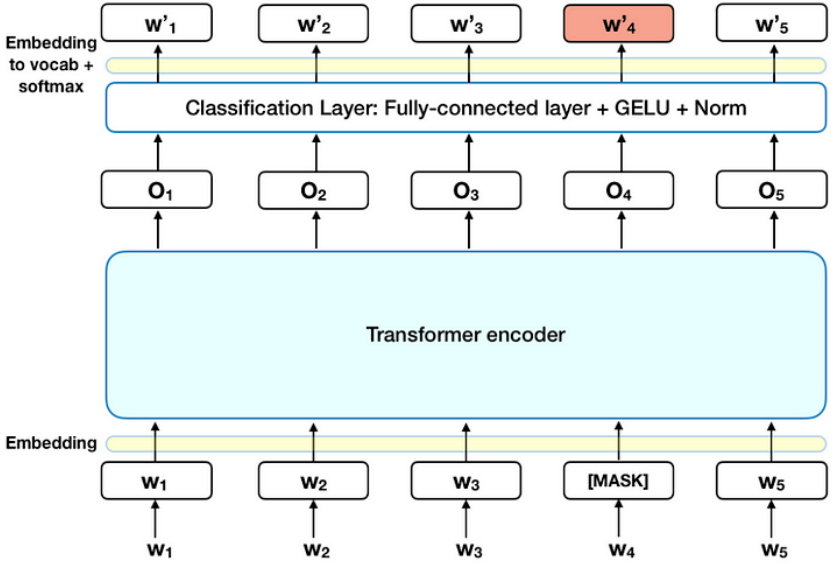
TD-IDF Vectorizer

	blue	bright	sky	sun
Doc1	0.707107	0.000000	0.707107	0.000000
Doc2	0.000000	0.707107	0.000000	0.707107

BERT Model

A word semantic based model

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.
ARA Vec for Arabic part



03

Similarity Computation

Chapter Wise

كتاب صلاة التراويح . 65
كتاب الجمعة . 66
كتاب الجزية والموادعة . 67
كتاب الإجارة . 68
كتاب اللباس . 69
كتاب أخبار الآحاد . 70
كتاب أحاديث الأنبياء . 71
كتاب التوحيد . 72
كتاب الوتر . 73
كتاب المزارعة . 74
كتاب الشفعة . 75
كتاب المظالم . 76
كتاب المرضى . 77
كتاب التقصير . 78
كتاب الوضوء . 79
كتاب المحصر . 80
كتاب الكفالة . 81
كتاب المساقاة . 82
كتاب الـ . 83

كتاب العيدين . 85
كتاب الاعتكاف . 86
كتاب الحوالات . 87
كتاب فضائل المدينة . 88
كتاب الأذان . 89
كتاب مناقب الأنصار . 90
كتاب الإيمان . 91
كتاب الإكراه . 92
كتاب الاستئذان . 93
كتاب فرض الخمس . 94
كتاب الزكاة . 95
كتاب الجنائز . 96
كتاب الهبة وفضلها والتحريض عليها . 97
كتاب الأضاحي . 98
كتاب الطلاق . 99
كتاب الدعوات . 100
كتاب فضل الصلاة في مسجد مكة والمدينة . 101
كتاب السلم . 102
كتاب الشركة . 103

Distinct Strings:
كتاب التمني . 1
كتاب فضائل القرآن . 2
كتاب الرهن . 3
كتاب الكسوف . 4
كتاب الغسل . 5
كتاب كفارات الأيمان . 6
كتاب الوصايا . 7
كتاب الطب . 8
كتاب التعبير . 9
استتابة المرتدين والمعاندين وقتالهم
كتاب الشروط . 11
كتاب الحيل . 12
كتاب القدر . 13
كتاب سجود القرآن . 14
كتاب العتق . 15
كتاب الأيمان والنذور . 16
كتاب الرقاق . 17

Whole Data

Mukarrarat

Similar Ahadis Data taken from GitHub

04

Equivalence Class

Equivalence class Analysis

Identical : 0.9888-1

Similar : 0.8- 0.988

Not Similar: < 0.8