Détection de Paraphrases

Proposition de Projet Final

Etienne Collin | 20237904,
Guillaume Genois | 20248507

Fondements Théoriques en Science des Données - STT3795

Section A

Stefan Horoi & Guillaume Huguet

Université de Montréal

Hiver 2024
À remettre le 13 Février 2024 à 23:59

Table des matières

Table des matières **1**

1 Objectif **2**

2 Ressources **2**

3 Livraisons Attendues **2**

4 Description **3**

5 Contributions **4**

1 Objectif

L'objectif principal de ce projet est de développer un système de classification de texte capable de déterminer si deux phrases données sont des paraphrases l'une de l'autre. Nous nous concentrerons sur l'utilisation d'un modèle Naive Bayes pour cette tâche, en explorant également la possibilité de générer des paraphrases si le temps le permet. En outre, nous évaluerons la généralité de notre modèle en le testant sur le dataset GLUE.

2 Ressources

— Le corpus MRPC pour l'entraînement et l'évaluation du modèle.

— Dataset GLUE pour tester la généralité du modèle.

— Outils de traitement du langage naturel (NLP) pour le prétraitement des données.

— Bibliothèques de machine learning en Python telles que scikit-learn pour la mise en œuvre du modèle Naive Bayes.

3 Livraisons Attendues

1. Code source du modèle Naive Bayes pour la classification de paraphrases.
2. Rapport détaillé sur l'approche utilisée, les résultats obtenus, et les éventuelles expérimentations sur la génération de paraphrases.
3. Évaluation des performances du modèle sur le dataset GLUE avec une analyse approfondie des résultats.
4. Présentation des conclusions et des pistes d'amélioration pour de futures recherches.

4 Description

Le projet se basera sur le corpus MRPC (Microsoft Research Paraphrase Corpus), qui propose une collection de paires de phrases annotées pour la détection de paraphrases. L'idée est d'implémenter un classificateur Naive Bayes pour évaluer la similarité sémantique entre les paires de phrases.

1. **Classification de Paraphrases :** Utiliser le modèle Naive Bayes pour la classification des paires de phrases en paraphrases ou non-paraphrases. - Expérimenter avec différentes représentations de texte telles que bag-of-words, TF-IDF, ou d'autres caractéristiques pertinentes pour le modèle Naive Bayes.
2. **Évaluation sur le Dataset GLUE (STS-B et QQP) :**
   - Tester la généralité du modèle en l'appliquant au dataset GLUE, qui propose une diversité de tâches de compréhension du langage naturel.
   - Analyser les performances du modèle sur différentes tâches du GLUE et tirer des conclusions sur sa capacité à généraliser.
3. **Approfondissements (si le temps le permet)**
   - Comparaison avec une implémentation SVM
   - Comparaison avec une implémentation random forest
   - Optimisation des performances
   - Génération de Paraphrases
     - Si le temps le permet, explorer des méthodes de génération de paraphrases. Cela pourrait impliquer l'utilisation de techniques de réécriture automatique ou de modèles de génération de texte pour créer des variations sémantiques des phrases.

5 Contributions

Notre équipe de deux personnes collaborera étroitement tout au long du projet en adoptant une approche de "pair-programming" et en travaillant conjointement sur les différentes facettes du problème. Étant donné que nous sommes confrontés à un domaine relativement nouveau pour nous, nous prévoyons explorer collectivement les étapes clés du projet.

Bien que nous prévoyions une répartition flexible des tâches, la nature collaborative du projet signifie que chaque membre de l'équipe participera activement à chaque phase, partageant les idées, débattant des approches, et contribuant à la prise de décisions. Cela permettra une compréhension approfondie du problème et une expertise partagée au sein de l'équipe.

1. **Exploration du Problème (Collaboratif)**
2. **Prétraitement des Données (Pair-Programming)**
3. **Implémentation du Modèle Naive Bayes (Pair-Programming)**
4. **Évaluation sur le Dataset GLUE (Pair-Programming)**
5. **Aprrofondissements :**
   - Comparaison avec une implémentation SVM (Guillaume)
   - Comparaison avec une implémentation random forest (Guillaume)
   - Optimisation des performances (Etienne)
   - Génération de Paraphrases (Etienne)
6. **Rapport (Collaboratif)**