INFORMATION RETRIEVAL

Flavio Forenza
flavio.forenza@studenti.unimi.it

Department of Computer Science,
University of Milan, Italy

September 10, 2021

INTRODUCTION

Nowadays, searching the web for information appears to be one of the simplest operations to perform. The difficulty perceived by the user in formulating a query has been gradually reduced by techniques capable of guiding his writing towards a correct generation of a query. These techniques allow to improve the performance of information search systems.

GOAL

The purpose of this project is to be able to experiment with the use of one of the most famous techniques, already present at the state of the art, able to "assist" the user in formulating a correct query: **Language Modelling**. Query expansion can be done using this concept to return a corpus of relevant documents.

DATASET DESCRIPTION

The dataset used for the experiments is the famous *Recipes1M+* [1], a collection created by MIT, consisting of more than one million culinary recipes. Of all these recipes, only a subset of 51235 documents of it was used due to their informative content which best fits the purpose of this study. The information about the line distributions for each recipe indicates that the instruction field contains a higher number than the information contained in the ingredients field.

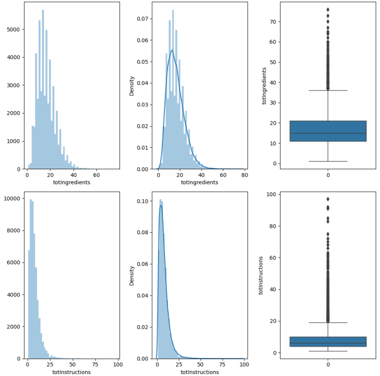

Figure 1. Distributions of lines per ingredients and instructions.

[1] J.Marin, A.Biswas, F.Ofli, N.Hynes, A.Salvador, Y.Aytar, I.Weber and A.Torralba, "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images", IEEE Trans. Pattern Anal. Mach. Intell., 2019

RANKING GENERATION

RANKING EVALUATION

LANGUAGE MODELS

SMOOTHING METHODS

CORE

TERM-TERM MATRIX

POSITIVE POINTWISE MUTUAL INFORMATIONS (PPMI)

SINGULAR VALUE DECOMPOSITION (SVD)

QUERY EXPANSION

PERPLEXITY

SYSTEM EVALUATION WITH DIFFERENT PARSERS

CONCLUSIONS