Games classifier

Team name

Members: Julia Cygan, Borys Adamiak, Patryk Flama
Supervisor: Marek Adamczyk

UWr

18 stycznia 2025

Your task

Goal and motivation

We want to be able to automatically assign tags (or genres) to games, based on their text description. Solution to such problem has real-world applications, such as game grouping/filtering or finding simmilar games or trends analysis. Additionally, in aspect of ML project, we want to make a small comparison of different models and data processing methods for such multilabel classification problem.

Your task

Goal and motivation

We want to be able to automatically assign tags (or genres) to games, based on their text description. Solution to such problem has real-world applications, such as game grouping/filtering or finding simmilar games or trends analysis. Additionally, in aspect of ML project, we want to make a small comparison of different models and data processing methods for such multilabel classification problem.

Info about the data

Steam has its own official API, from which we want to download all of the data (and since Steam is the largest library it will allow for a lot of diverse, high quality, data). Currently there are above 100'000 games, which does create a large dataset. After cleaning the data (removing empty descriptions or tags, removing tags that occur only once) we ended up with a dataset of size around 50'000 games and 400 tags or 100 genres.

Methods

Data processing

- Bag of Words - binary vector records if word appears in text (input representation) ✓
- TF-IDF - term frequency * inverse document frequency (input representation) ✓
- Hashing vectorizer - method to generate low-dimensional input representation (input representation) ✓
- multi label binary vector (output representation) ✓

Models

- KNN ✓
- Logistic Regression ✓
- Decision Trees + Random Forest ✓
- Naive Bayes ✓
- Simple perceptron-based neural network ✓
- Support Vector Machine ✓

Evaluation

After analyzing the problem we came to conclusion that evaluaion methods are very interesing and important part of this project.

Evaluation

After analyzing the problem we came to conclusion that evaluaion methods are very interesing and important part of this project.

- We dont want to falsely assign a tag to a game that should not have it
- Its more important to assing high percentage of tags to games, than to assign as many as possible
  - Game should have 10 tags, but we only assign 8 (not bad)
  - Game should have 1 tag, but we do not assign any (this is worse)

Evaluation

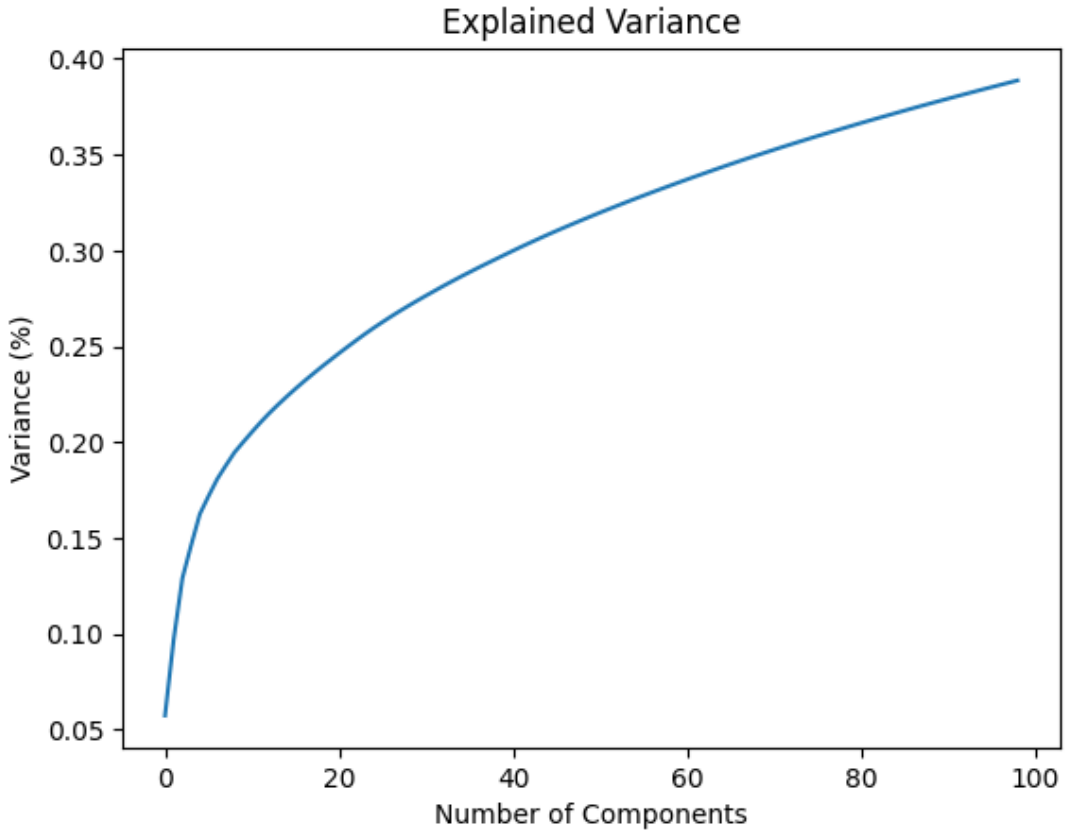
- Recall *TP/(TP+FN)* - we prefer to have more FN than to have an TP ✓
- F1-score *(2 * precision * recall) / (precision + recall)* - nice name, but also it combines precision with recall thus both TP and FN are equally expensive ✓
- Hammming loss
- Intersection over union score
- Exact match

Input data

First we tried some unsupervised methods to check if we can find some patterns in the data

Input data

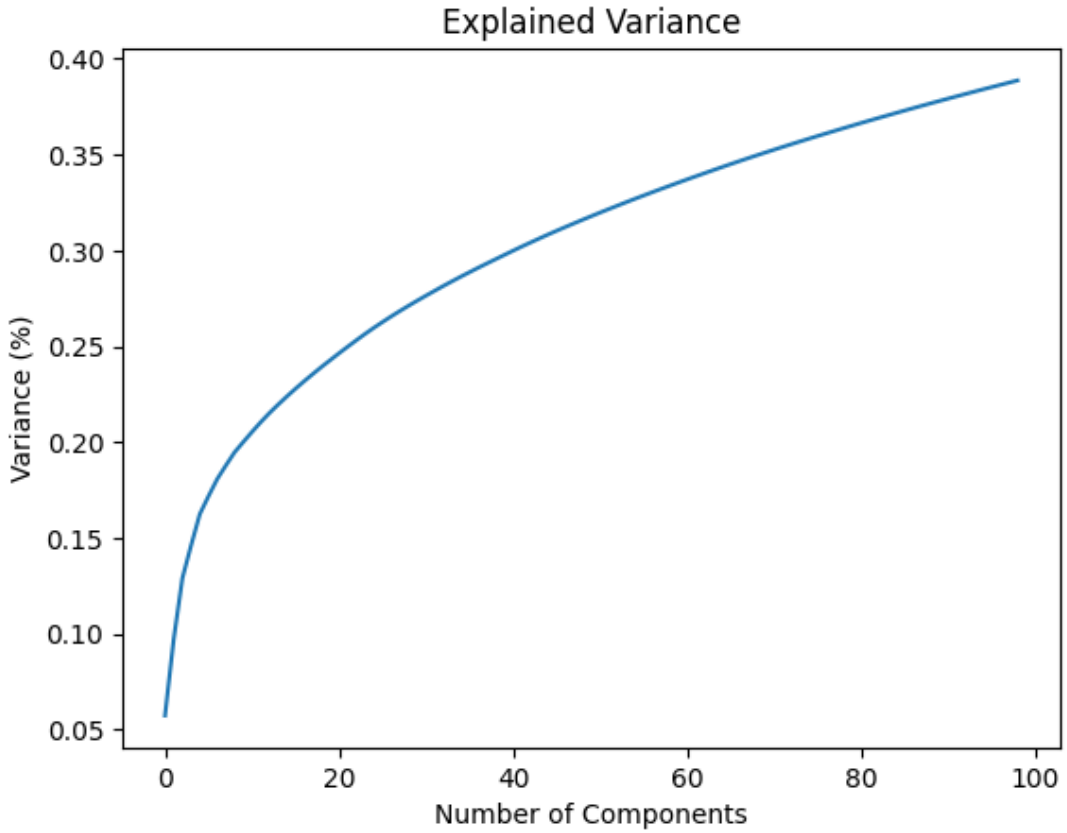
PCA on Bag of Words representation of the data (to 10'000 words)



Nice, out of 10'000 dimensions we can create 100 that 'explain' about half of the data.

Input data

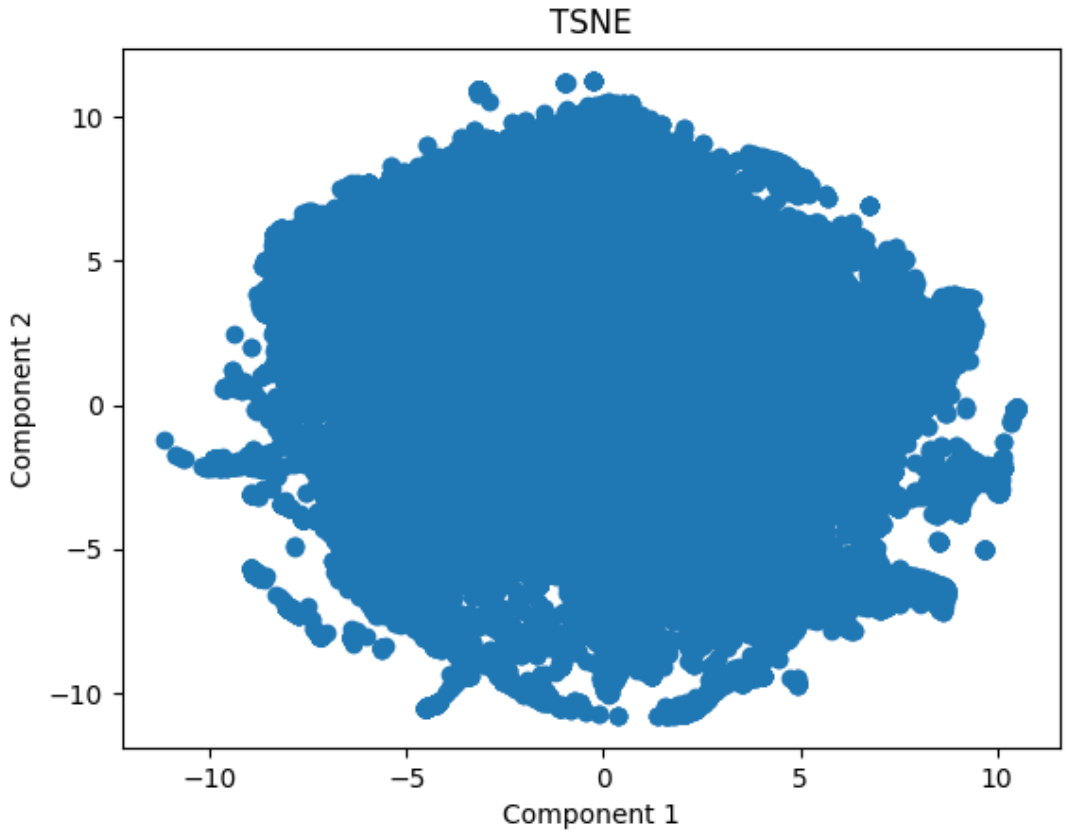
PCA on Bag of Words representation of the data (to 10'000 words)



Nice, out of 10'000 dimensions we can create 100 that 'explain' about half of the data.

Input data

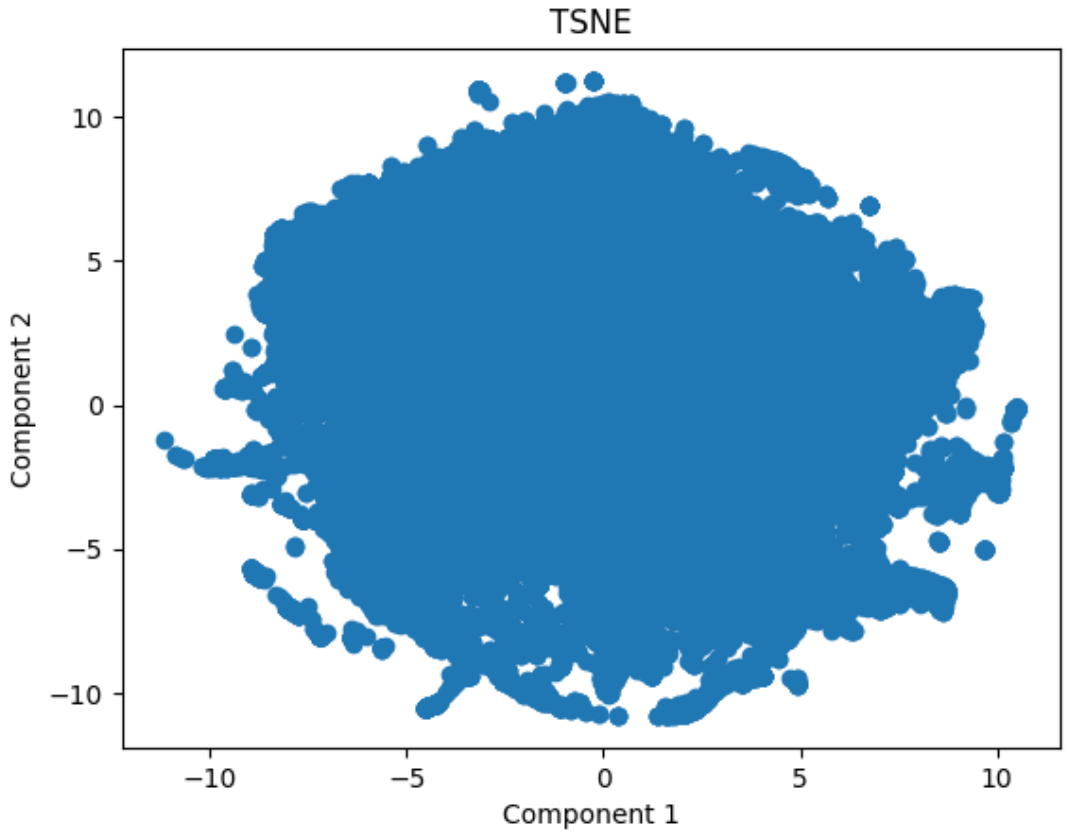
t-SNE representation of same data



And it does not look that helpful :<

Input data

t-SNE representation of same data



And it does not look that helpful :<
But what about combining PCA with t-SNE?

Input data

But what about combining PCA with t-SNE?