

CLASSIFICATION OF AMAZON REVIEWS BY TOPIC



Text mining and Search

Academic Year 2019/2020

PROJECT WORKFLOW

THE TASK

Text Classification of 14000 Amazon Reviews in 7 topics:



DIGITAL MUSIC
2000 Reviews



ALL BEAUTY
2000 Reviews



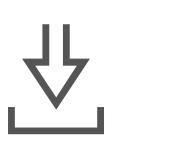
GIFT CARDS
2000 Reviews



LUXURY BEAUTY
2000 Reviews



APPLIANCES
2000 Reviews



SOFTWARE
2000 Reviews



MAGAZINE SUBSCRIPTION
2000 Reviews

The image displays a collection of five Amazon product review snippets arranged in a grid, followed by a mobile phone screen showing a summary of these reviews.

Review 1 (Top Left): Angelo Monachesi, ★★★★★ Acquisto verificato. Revisionato in Italia il 24 novembre 2019. Formato: Copertina flessibile. **Summary:** "Approach to machine learning... book because totally new on this... perfect for every software engineer without machine learning background". [Visualizza altro](#)

Review 2 (Top Middle): Amazon Customer, ★★★★★ Acquisto verificato. Revisionato nel Regno Unito il 30 agosto 2019. **Summary:** "Short and concise... For the most part, I liked the short and concise explanations. They were so concise I found my self reading and rereading sentences simply because there was so much information condensed into them. I disliked the treatment of backpropagation, which was almost non-existent...". [Visualizza altro](#)

Review 3 (Top Right): Andrei Damian, ★★★★★ Acquisto verificato. Revisionato nel Regno Unito il 6 febbraio 2019. **Summary:** "Perfect for every software engineer without machine learning background... Andriy Burkov's 'The Hundred-Page Machine Learning Book' became a mandatory reading for software engineers in our Data Science company the second day I received my copy. Although not a book for machine learning engineers and scientists with formal education in the field or...". [Visualizza altro](#)

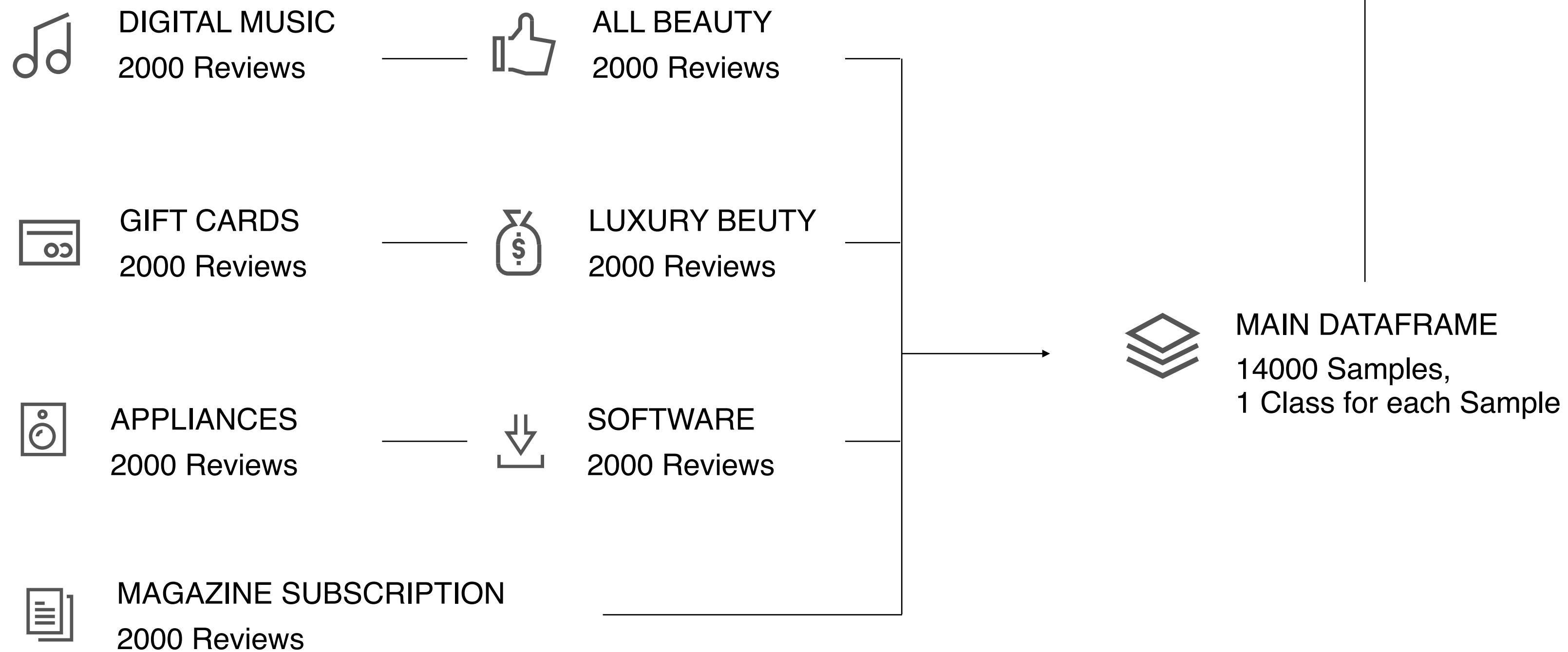
Review 4 (Bottom Left): Olena Karazieva, ★★★★★ Acquisto verificato. Revisionato nel Regno Unito il 22 settembre 2019. **Summary:** "Awesome Book about Machine Learning... I have read this book and coming back to it again and again. The subj is really well-explained and illustrated by examples. I tried other books and the articles on internet but Burkov's book beats all of them. If you want understand what ML is, this is definitely the book...". [Visualizza altro](#)

Review 5 (Bottom Middle): jamo, ★★★★★ Acquisto verificato. Revisionato nel Regno Unito il 29 dicembre 2019. **Summary:** "One of the books to check to learn ML... How many pages can have a book whose purpose is to explain, with a not trivial level of detail, what Machine Learning (ML) is? It's natural to think of games, like Lords of the Rings, like... Let's think about it: ML is...". [Visualizza altro](#)

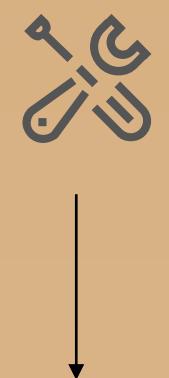
Mobile Phone Screen: Shows the 'Recensioni clienti' (Customer Reviews) section of the Amazon website. It displays a summary: 244 reviews, 4,3 su 5 stelle (4.3 stars), and a distribution bar chart for star ratings (63% for 5 stars, 18% for 4 stars, 9% for 3 stars, 4% for 2 stars, 6% for 1 star). Below this, it shows a section titled 'Immagini del cliente' (Customer Images) featuring small thumbnail images of products. At the bottom, it lists reviews for Andrei Damian, matching the ones shown above.

PROJECT WORKFLOW

THE PIPELINE



PREPROCESSING



TEXT REPRESENTATION



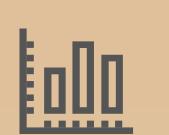
DIMENSIONALITY REDUCTION

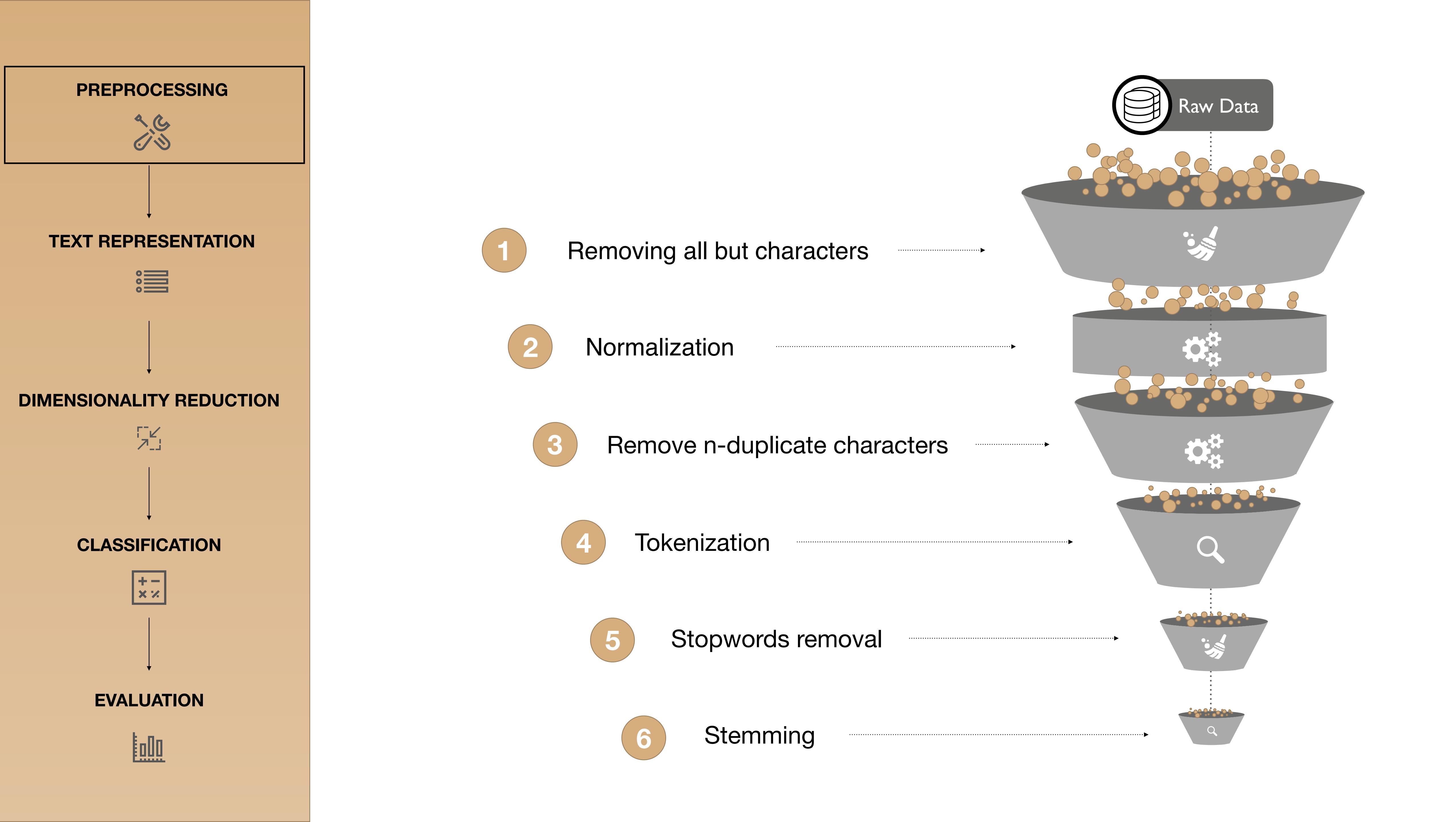


CLASSIFICATION

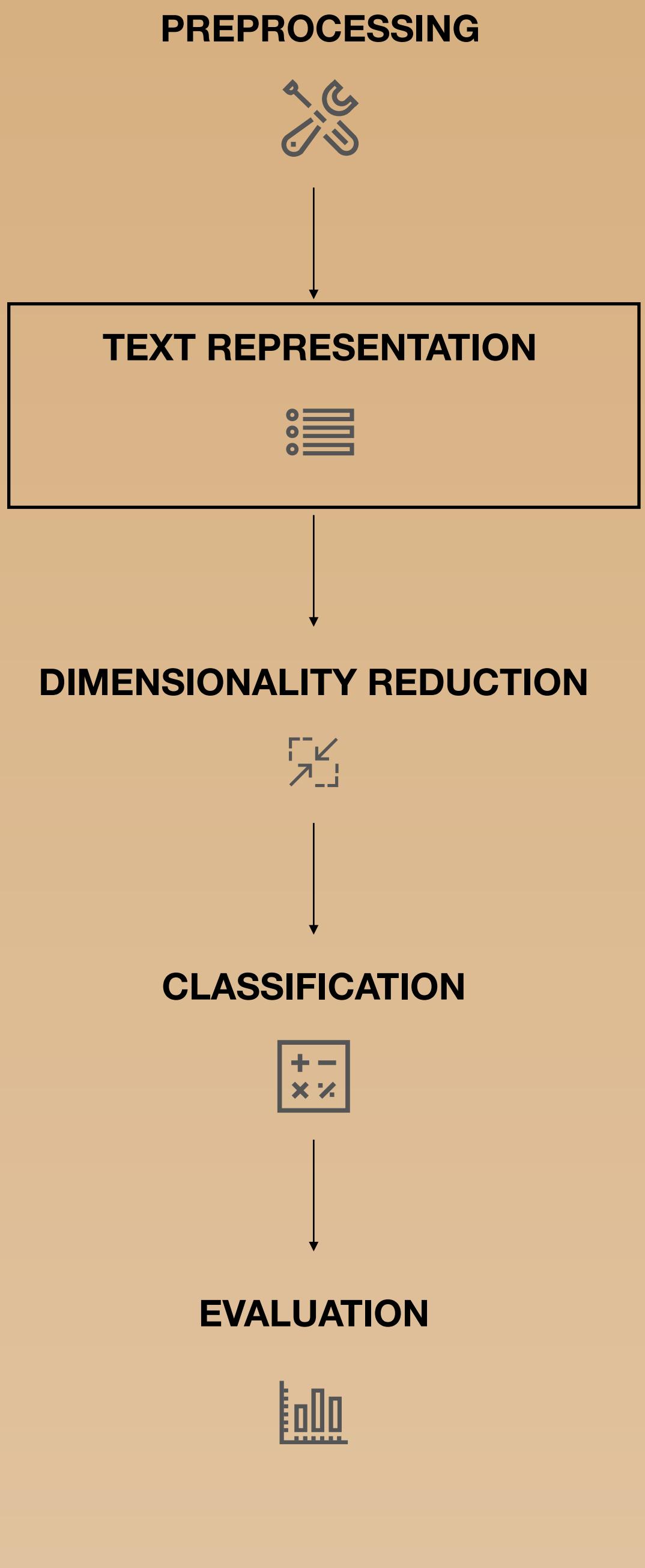


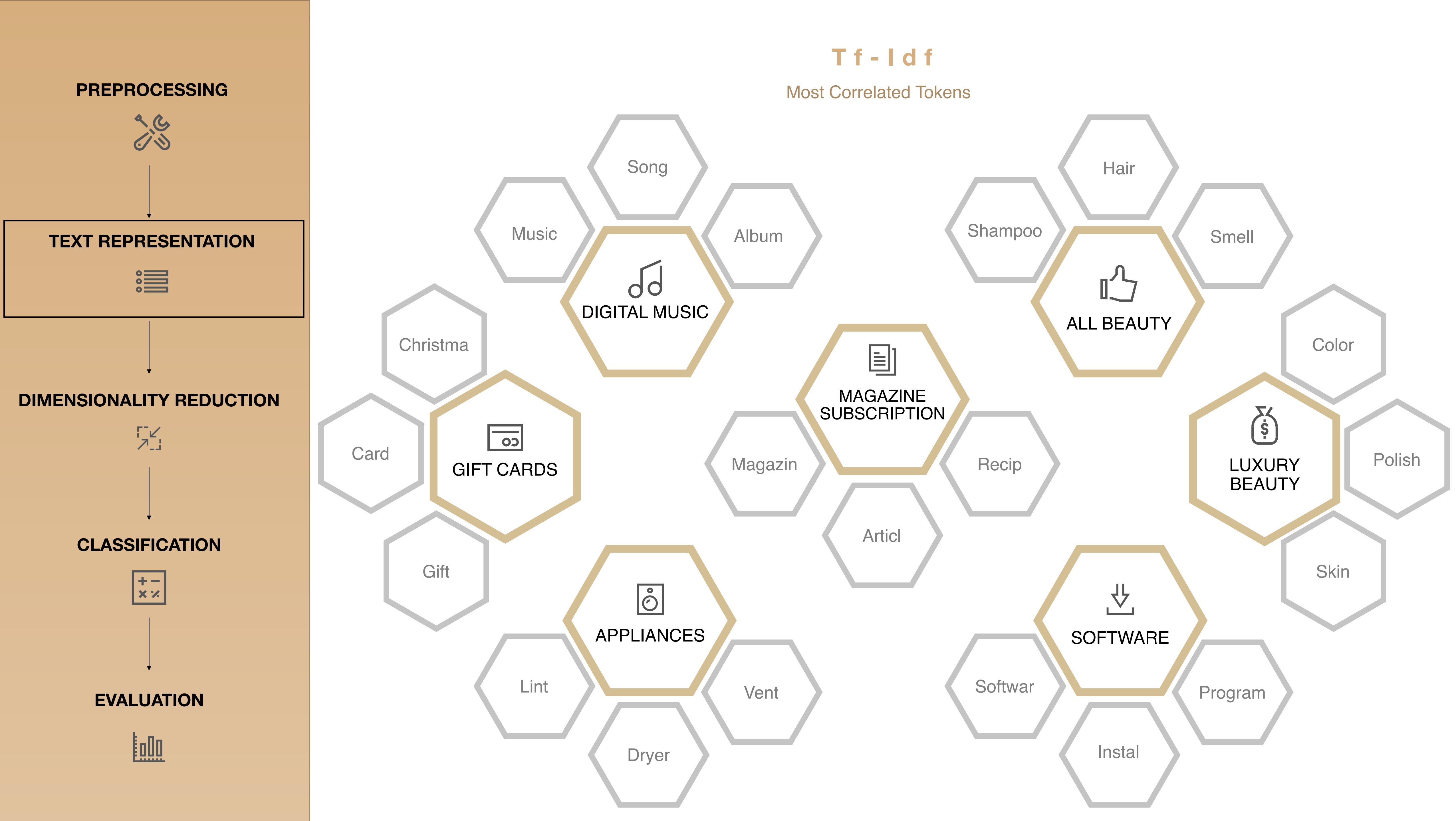
EVALUATION



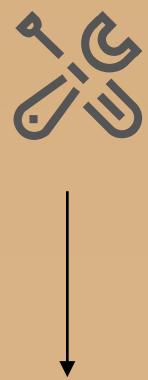


Bag of Words





PREPROCESSING



TEXT REPRESENTATION



DIMENSIONALITY REDUCTION



CLASSIFICATION



EVALUATION

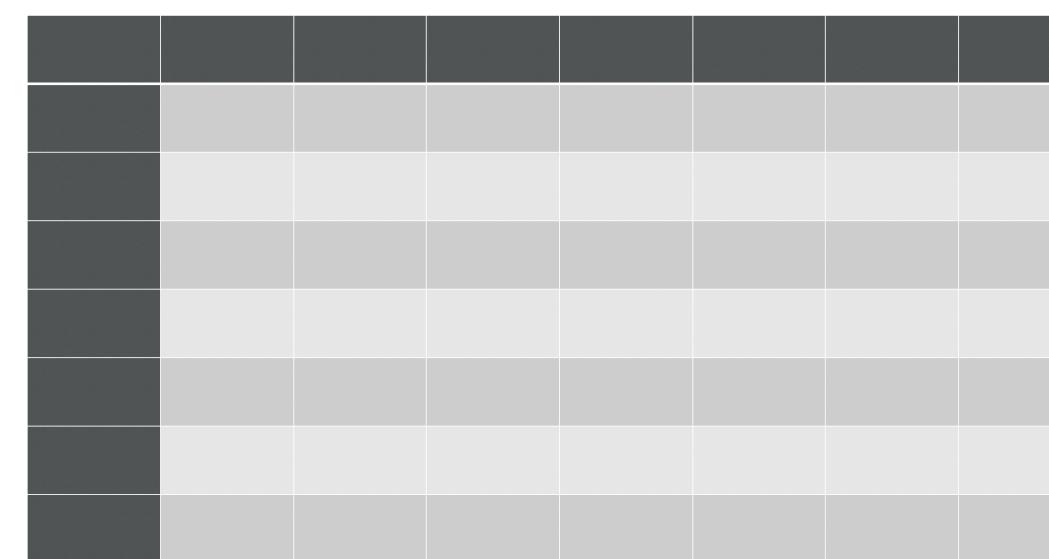


Features Selection

Maintaining only the
best 75 % features
according to χ^2
independence test

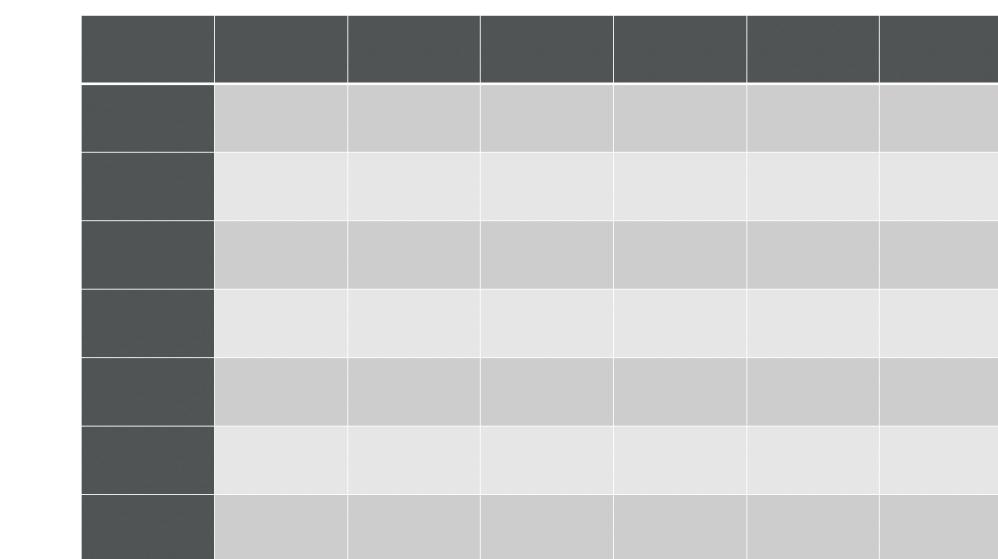


Tf-Idf Matrix



14000 x 4852

Tf-Idf Matrix Reduced



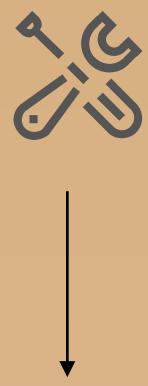
14000 x 3639



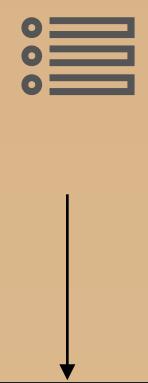
Execution Time:
1 second

Features Synthetization

PREPROCESSING



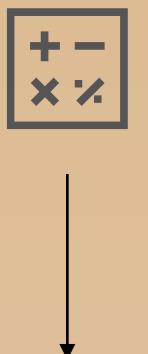
TEXT REPRESENTATION



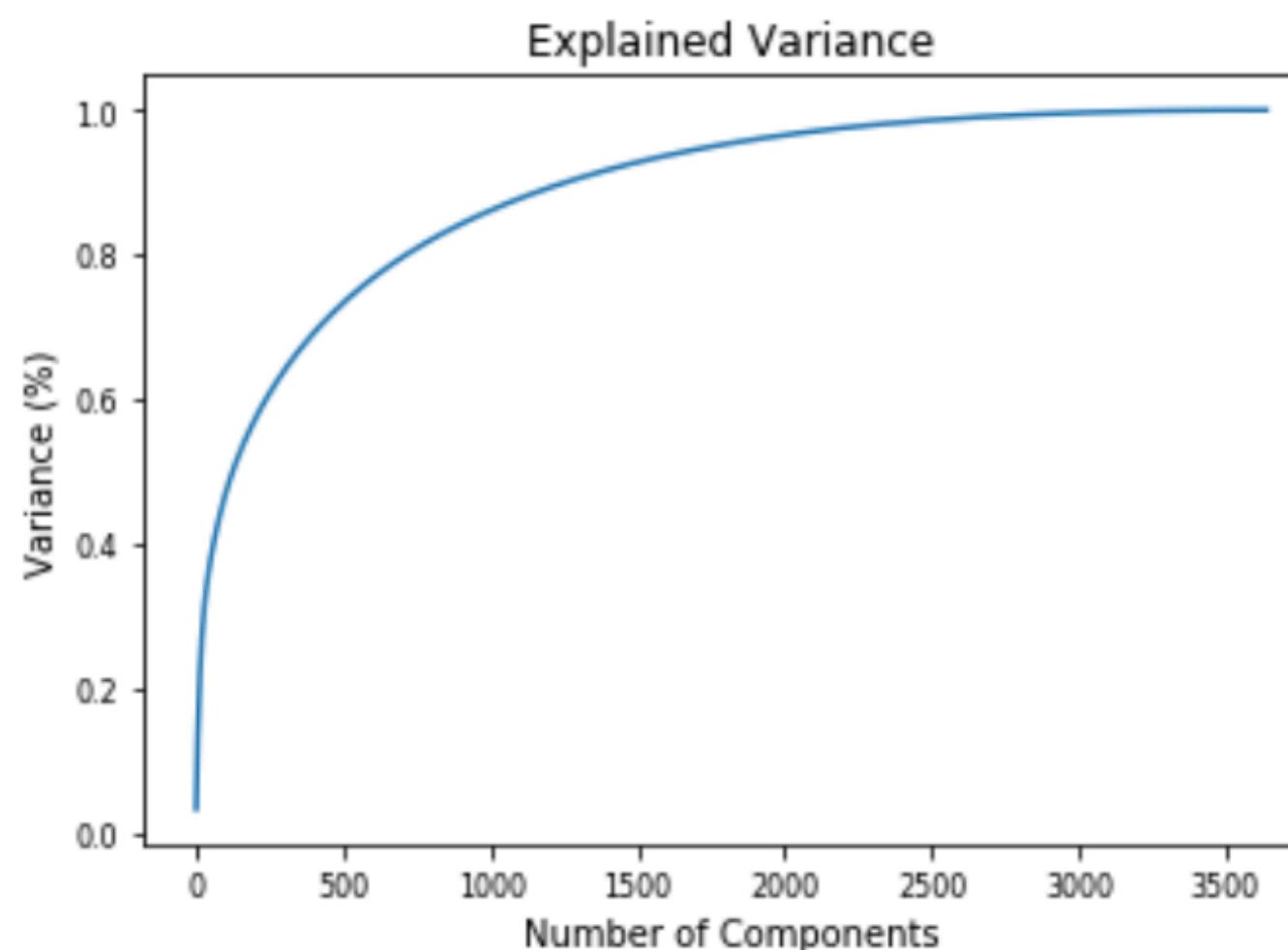
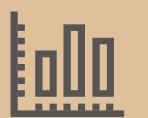
DIMENSIONALITY REDUCTION



CLASSIFICATION



EVALUATION



Maintaining the first 2000
components that can explain more
than 90% of Variance

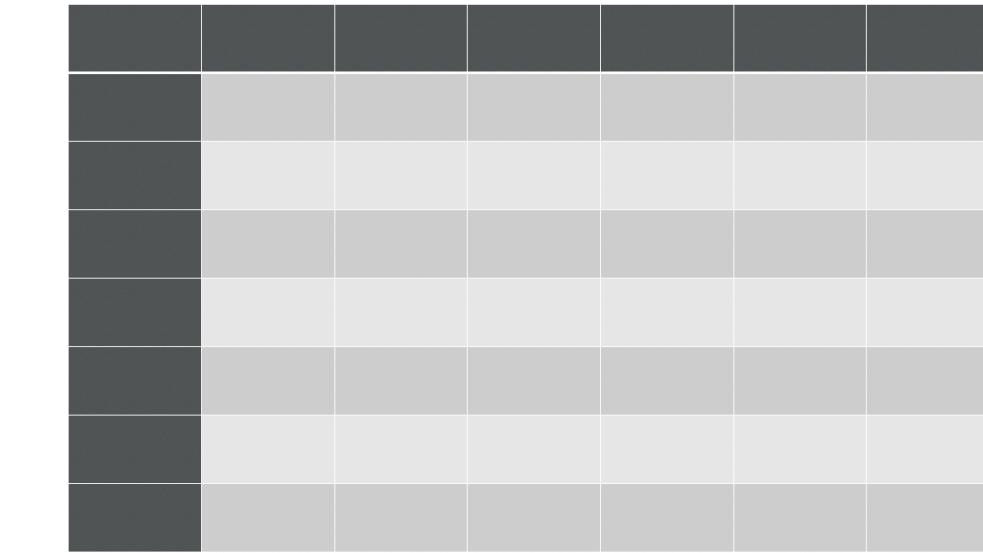


Execution Time:

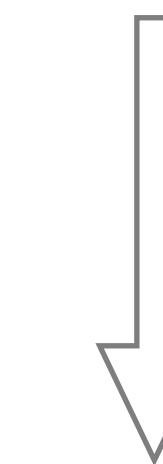
76 seconds (fit)

117 seconds (fit-transform)

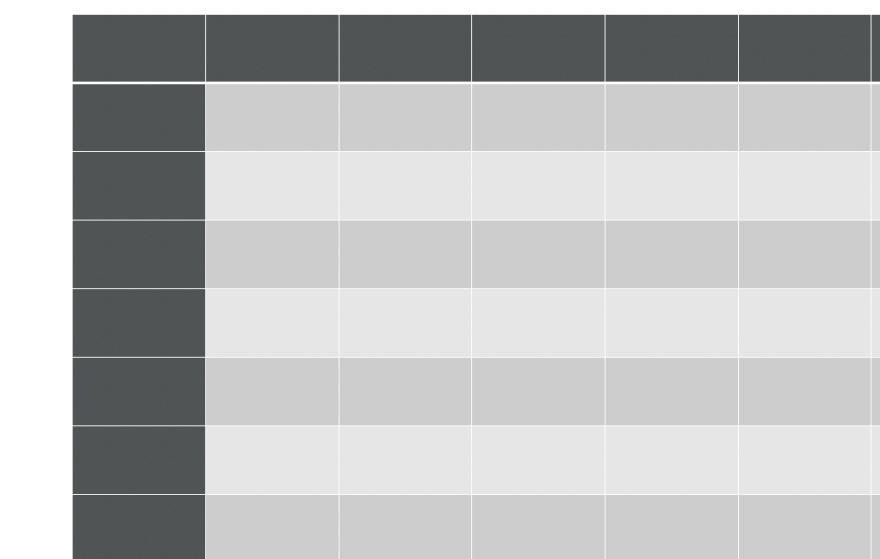
Tf-Idf Matrix Reduced



14000 x 3639



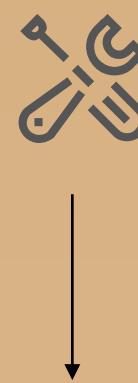
PCA Matrix



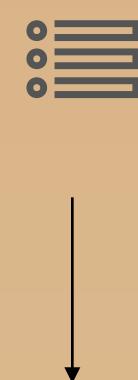
14000 x 2000

5 - f o l d C r o s s V a l i d a t i o n

P R E P R O C E S S I N G



T E X T R E P R E S E N T A T I O N



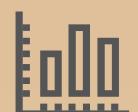
D I M E N S I O N A L I T Y R E D U C T I O N



C L A S S I F I C A T I O N



E V A L U A T I O N



1

SVC

Support Vector Classifier

Penalty: Norm L2
Loss: Squared Hinge
C: 1.0



Training Time:
46 seconds



Training Costs:
239 seconds

2

LR

Logistic Regression

Penalty: Norm L2
Solver: lbfsgs
C: 1.0



Training Time:
65 seconds



Training Costs:
258 seconds

3

MLP

MultiLayer Perceptron

H L sizes: 100,50,50,50
Activation: ReLU
Solver: AdaM



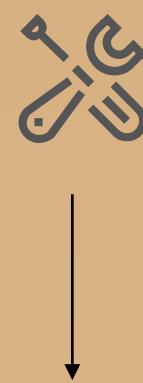
Training Time:
336 seconds



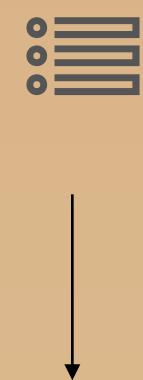
Training Costs:
529 seconds

5 - f o l d C r o s s V a l i d a t i o n

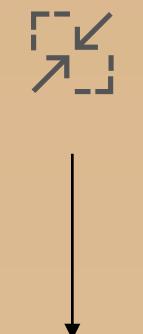
PREPROCESSING



TEXT REPRESENTATION



DIMENSIONALITY REDUCTION



CLASSIFICATION



EVALUATION



Training Set: 12600 examples (90%)

Test Set: 1400 examples (10%)

1

SVC

Support Vector Classifier

Mean Accuracy:

0.865397

Standard Deviation:

0.004690

2

LR

Logistic Regression

Mean Accuracy:

0.859206

Standard Deviation:

0.008415

3

MLP

MultiLayer Perceptron

Mean Accuracy:

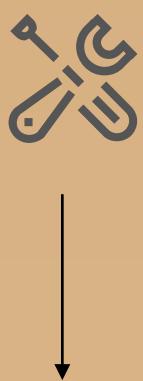
0.833651

Standard Deviation:

0.012083

Test set predictions

PREPROCESSING



TEXT REPRESENTATION



DIMENSIONALITY REDUCTION



CLASSIFICATION



EVALUATION



Training Set: 12600 examples (90%)

Test Set: 1400 examples (10%)

1 SVC

Support Vector Classifier

Mean Accuracy:

0.865397

Standard Deviation:

0.004690

	precision	recall	f1-score	support
Digital_Music	0.81	0.79	0.80	199
All_Beauty	1.00	0.97	0.98	183
Software	0.86	0.87	0.86	202
Gift_Cards	0.75	0.89	0.81	210
Luxury_Beauty	0.84	0.82	0.83	204
Appliances	0.91	0.88	0.89	214
Magazine_sub	0.97	0.88	0.92	188
accuracy			0.87	1400
macro avg	0.88	0.87	0.87	1400
weighted avg	0.87	0.87	0.87	1400

Classification Costs:
1 second

CLASSIFICATION OF AMAZON REVIEWS BY TOPIC

CONCLUSIONS

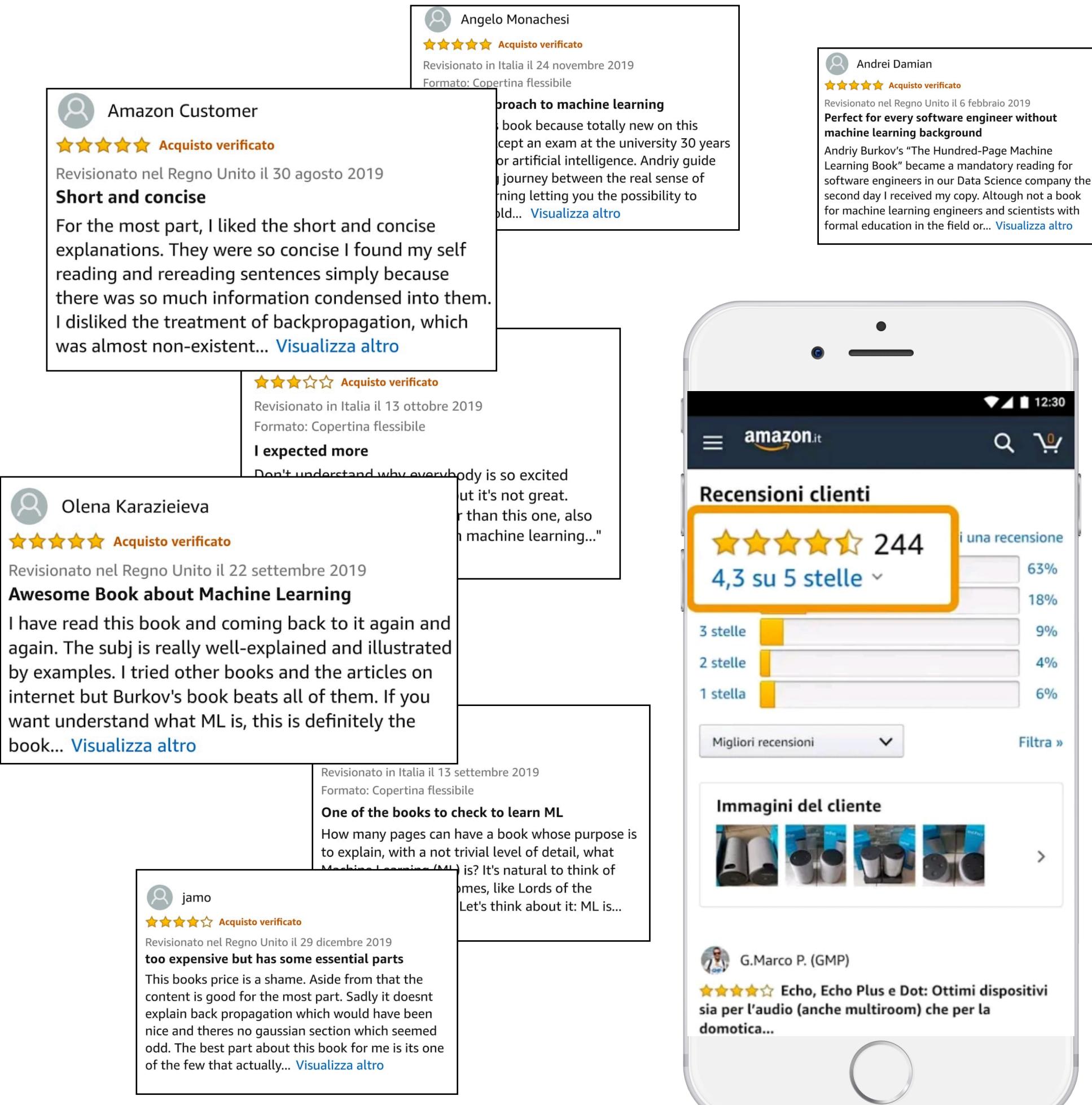
- ML algorithms are able to correctly classify Amazon Reviews by topic.

ISSUES

- High Computational Costs in features synthetization;
- Limited number of observation for each category due to prohibitive computational costs.

NEXT STEPS

- Scale the problem in dimension with less expensive dimensionality techniques, such as LDA.



THANKS FOR YOUR ATTENTION



Raffaele Anselmo - 846842

Lorenzo Pastore - 847212