

List of projects

Each project has minimum and desirable requirements and, according to minimum requirements, each project has its own difficulty:

- A - easy
- B - moderate
- C - hard

It means that even for A-project you may obtain good marks. You may do some of the desirable requirements, you may go beyond or propose other solutions - it's up to you. The more you do the better. You are free to apply methods described in other exercises. These projects are oriented on your ability to work with text, visual data, and trends and on discovering marketing insights from data. In case of any difficulty you can reach me via e-mail

yaroslav.kozyrev@ipsos.com

For these projects I suggest to use Google Colab, Spark, Cloud solutions or various APIs (e.g. to extract data) but if you decide not to use them or if it's not your case - do not worry, it will not affect your final mark.

Dataset of your choice - you may use any opensource dataset related to your project.

EDA – exploratory data analysis.

For data visualization you may use matplotlib, plotly, bokeh, seaborn and other packages.

Output format:

- A link to a Git repository with all scripts and comments so that it's easy to replicate
- OR
- Well-organized Jupyter notebook with comments and .zip archive with data

NLP

1. **(B+) Text deduplication.** With a dataset of your choice, using unsupervised learning techniques, identify duplicates or duplicated parts of a corpus (e.g. “Check out my new album on Spotify” VS “Check out my new album on Soundcloud”)

Minimum requirements: EDA; identification of (near-)duplicates; deduplication of common segments; construction of a topic model of your choice; name some topics; visualization (e.g. word clouds, dendrogram, bubble chart of PoS tags, etc); provide and explain marketing/research insights from outputs

Desirable: Use BERT-like sentence embeddings; do comprehensive text cleaning; output top-10 most frequent words and bigrams per topic in case of clustering method; output topic probabilities per post in case of LDA; Do parameter selection and explain; Name obtained topics; EDA of your outputs – topic sizing, charts, statistics, etc.

Hint: Use text segmentation techniques (sentence or clause level); don't take too big dataset (<5k is enough)

2. **(A) Topic modeling.** With a dataset of your choice, perform different approaches of topic modeling - LDA and Clustering - and provide marketing insights.

Minimum requirements: EDA; topic model with LDA; topic model with clustering techniques; name topics; visualization (e.g. word clouds, dendrogram, bubble chart of PoS tags, etc); comparison of outputs; EDA on results; provide and explain marketing/research insights from outputs

Desirable: Use 2 or more languages in a dataset; Usage of sentence embeddings; output top-10 most frequent/relevant* words and bigrams per topic for clustering method; output topic probabilities per post with LDA; Usage of UMAP and HDBScan libraries and its comparison with other methods; Do parameter selection and explain;

3. **(C) Classification.** You will work with two datasets of consumer reviews ([for example](#)). With one labeled sentiment dataset of your choice, train a sentiment classification model. Project it on the second dataset. Perform LDA and clustering topic models on the second dataset, compare results and provide marketing/research insights.

Minimum requirements: Exploratory data analysis; sentiment model with TensorFlow/PyTorch/SpaCy/Transformers frameworks; output top-10 most relevant words/bigrams per topic in case of clustering method; output topic probabilities per post in case of LDA; name some topics; calculate average sentiment per topic; visualization (e.g. word clouds, dendrogram, bubble charts of PoS tags, etc); provide and explain marketing insights from outputs.

Desirable: Use 2 or more languages in datasets; Use BERT-like models; Use HDBScan library; Parameter tuning

Computer Vision

1. **(B) Image classification.** With a dataset of food/drinks/brands/logos, create a multilabel image classification model using any pretrained model (e.g. VGG16, InceptionV3, EfficientDet/EfficientNet, etc.)



Minimum requirements: find/collect and, if needed, annotate enough data for at least 3 classes; use built-in data augmentation functions (e.g. [from tensorflow](#)); use transfer learning techniques to train a model; evaluate it on test data; provide and explain how can you apply this model in marketing/market research field to find insights from pictures

Desirable: Use [Albumentations](#) for image augmentation ([example](#)); make some photos and test your model on them; test your model on random pictures of landscapes/inside-outside environments/etc; provide possible marketing ideas of it.

2. **(C) Logo detection.** With any labeled logo dataset of your choice, create a multilabel object detection model using any pretrained weights and architectures.

Minimum requirements: use any labeled dataset (example [1](#), [2](#)) and convert annotations into required format for your model's input; use any augmentations; use transfer learning techniques to train a model; evaluate it on test data; provide and explain how your model can be used in marketing domain;

Desirable: include in your dataset 2 look-alike logos, e.g. Sun Microsystems and

Columbia.   Test your model on these logos; output 2nd best prediction

3. **(A) Image Clustering.** With any dataset containing at least 500 different pictures, do image clustering modeling and output generated captions with [this model](#).

Minimum requirements: use any picture dataset; get features with any pretrained model; do dimensionality reduction; run clustering algorithm; run image captioning model on your images; output to excel; provide and explain how your results can be used in marketing domain;

Desirable: visualize top-1 picture of your clusters on a 2D/3D scatterplot (points are pictures); output results of your clustered pictures to separate folders (e.g. Folder 1

contains pictures from Cluster 1, etc.); do text clustering on obtained captions and compare results;

Hint: to get top-1 relevant picture of a cluster – find a mean vector of pictures from this cluster, calculate distance between mean vector and pictures. The closest picture to mean vector will be your top-1.

Time Series

1. **(B) Clustering and prediction.** Collect at least 20 different trends (max. 100) of at least 3 different categories (max. 10) from Google Trends ([link 1](#), [link 2](#)). Minimum timeframe is 3 years. Cluster time series and make predictions.

Minimum requirements: explore time series: EDA, identify and handle outliers, check statistics, etc.; do clustering of a dataset; make predictions for each series and each cluster and compare results; explain obtained clusters and give names; data visualization of the original data and outputs; explain predictions and cause of outliers from market and global world point of view; provide marketing insights of your work

Desirable: use Facebook Prophet / Neural Prophet / Etna packages for Time Series modelling.