SCAN MY COFFEE

OCR TOOL FOR COFFEE RECOMMENDATIONS

BEN POH, DSI24

NOV 2021

CONTENT

- Background and Problem Statement
- Architecture Overview
- Part I: Optical Character Recognition (OCR) Tool
- Part II: Recommender System – (TFIDF, Auto-Encoder, Word2Vec, BERT)
- Areas for improvement
- Summary

BACKGROUND AND PROBLEM STATEMENT

- **Background**
  - Many OCR tools for text detection, designed for specific purposes
  - Popular apps in food-tech (e.g. Vivino), but nothing for coffee
  - Coffee consumers getting more sophisticated
  - Keen to know about origin, processing method, taste profile of bean

- **Problem Statement**

  Create a proof-of concept for an OCR tool that is optimised for single-origin coffee packaging;

  By using text data generated from the OCR tool, identify the ideal model that can most accurately recommend 5 coffees based on cosine similarity from an online store.

ARCHITECTURE OVERVIEW

PART 1: OCR

IMAGE
(300 PICS)

TEXT DETECTION
(EAST)

TEXT RECOGNITION
(TESSERACT)

(cleaning)

CORPUS

Dataset: 350 coffee descriptions scraped from SweetMarias.com

(string)

QUERY

['not, even, tanzania, friar, plums, molasses, washed']

(string)

MODELS

TF-IDF

- Basic, but fastest
- Fit for purpose: query only has key words

Auto-Encoder

- Feed-forward 6-layer
- Fair results, but tedious to train

Word2Vec

- Good at predicting association using neighbouring words (e.g. molasses sweetness)

BERT

- Captures semantics, BUT…
- Poor when 'irrelevant' words are in query

PART 2: RECOMMENDER SYSTEM

(350 vectors)

(one vector)

Cosine Distance

RECOMMENDATIONS

Top 5 from Sweet Marias dataset

OCR OVERVIEW

PART 1: OCR

IMAGE
(300 PICS)

Circular Text

Small Fonts

Angled Text

Poor Text Resolution

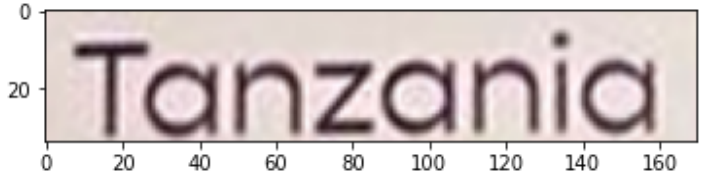
TEXT DETECTION
(EAST)

- Convolutional network trained on natural scene images (ImageNet)
- Confidence threshold at 80%, Non-Maximum Suppression at 40%
- **Two outputs:**
  1. Probability of whether area contains text
  2. Coordinates of bounding box

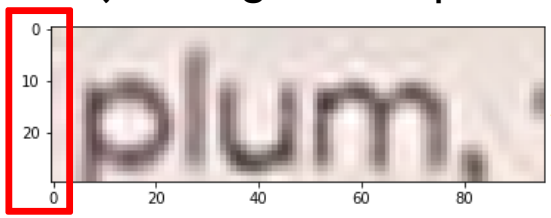
TEXT RECOGNITION
(TESSERACT)

- Optimisation by evaluating results after each run – what is important?
- Confidence threshold at 70%, padding at 5%
- Relaxed thresholds if <3 words captured
- **Output**: Text + probability score

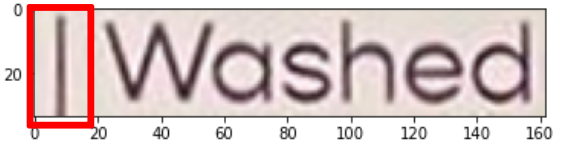
OCR EXAMPLE



Adjust height to 32 pixel

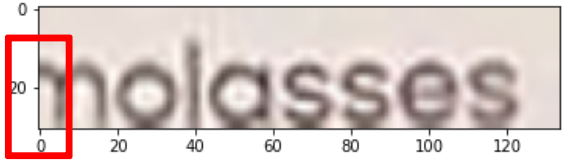


Drop non-words / low confidence

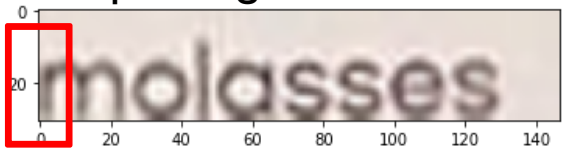


'| Washed' ? 'I Washed'? 'Washed'?

No padding



5% padding



RECOMMENDER OVERVIEW

PART 1: OCR

IMAGE
(300 PICS)

TEXT DETECTION
(EAST)

TEXT RECOGNITION
(TESSERACT)

(cleaning)

CORPUS

Dataset: 350 coffee descriptions scraped from SweetMarias.com

(string)

QUERY

['not, even, tanzania, friar, plums, molasses, washed']

(string)

MODELS

TF-IDF

- Basic, but fastest
- Fit for purpose: query only has key words

Auto-Encoder

- Feed-forward 6-layer
- Fair results, but tedious to train

Word2Vec

- Good at predicting association using neighbouring words (e.g. molasses sweetness)

BERT

- Captures semantics, BUT…
- Poor when 'irrelevant' words are in query

PART 2: RECOMMENDER SYSTEM

(350 vectors)

(one vector)

Cosine Distance

RECOMMENDATIONS

Top 5 from Sweet Marias dataset

CORPUS (DATASET)

OCR tool → must have real-world application

- ✓ Consumers: what else can they buy
- ✓ Retailers: sell what consumers want

CORPUS

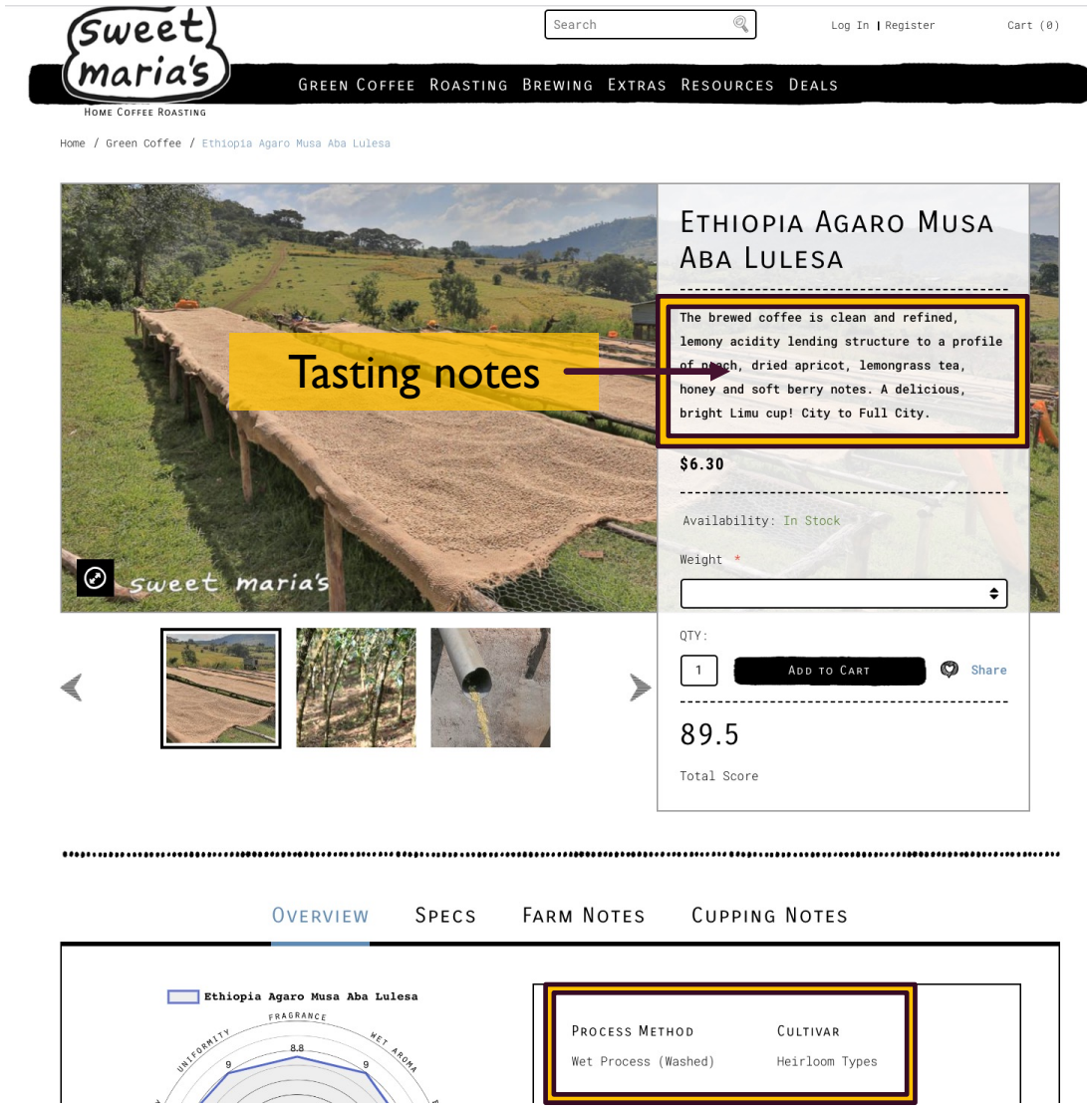
Dataset: 350 coffee descriptions scraped from SweetMarias.com

Pre-processing

- Tokenisation, lemmatisation, stop-words
- Applied on all 350 documents in the corpus (except BERT)

Cleaned Query

- Original: [‘not, even, tanzania, friar, plums, molasses, washed’]
- Cleaned: **[‘tanzania, friar, plum, molasses, washed’]**



RECOMMENDER OVERVIEW

PART 1: OCR

IMAGE
(300 PICS)

TEXT DETECTION
(EAST)

TEXT RECOGNITION
(TESSERACT)

CORPUS

Dataset: 350 coffee descriptions scraped from SweetMarias.com

(string)

(cleaning)

QUERY

['not, even, tanzania, friar, plums, molasses, washed']

(string)

MODELS

TF-IDF

- Basic, but fastest
- Fit for purpose: query only has key words

Auto-Encoder

- Feed-forward 6-layer
- Fair results, but tedious to train

Word2Vec

- Good at predicting association using neighbouring words (e.g. molasses sweetness)

BERT

- Captures semantics, BUT…
- Poor when 'irrelevant' words are in query

PART 2: RECOMMENDER SYSTEM

(350 vectors)

(one vector)

Cosine Distance

RECOMMENDATIONS

Top 5 from Sweet Marias dataset

MODEL OVERVIEW

TF-IDF
- Baseline

Auto-Encoder
- Unsupervised learning, trained on corpus
- Encode → Bottleneck → Decode (Machine Translation)
- Semantic information captured in word embeddings at Bottleneck

Word2Vec
- Learns semantic association by windows (preceding/succeeding words)
- Disadvantages: No vectors produced on unseen word / stores only one meaning per word

BERT
- Encode/Decode as basis, but uses 'attention mechanism'
- Bi-directional, words not treated as independent variables

MODELS

TF-IDF
- Basic, but fastest
- Fit for purpose: query only has key words

Auto-Encoder
- Feed-forward 6-layer
- Fair results, but tedious to train

Word2Vec
- Good at predicting association using neighbouring words (e.g. molasses sweetness)

BERT
- Captures semantics, BUT…
- Poor when 'irrelevant' words are in query

PART 2: RECOMMENDER SYSTEM

Corpus
(350 vectors)

RECOMMENDATIONS

Cosine Distance → **Top 5** from Sweet Marias dataset

Query
(one vector)

RECOMMENDER SYSTEM - RESULTS

Query from OCR:

[tanzania, molasses, plum, friar, washed]

Results:

- **TF-IDF**

Name	Description	Process
Colombia Tolima Productores de Ibagué	Unrefined sugar sweetness is central to the cup, accented by top notes of oatmeal cookie, molasses, dried date and cola nut, with plum-like acidity. City+ to Full City+.	Wet Washed
Colombia Ibagué Rio Combeima	Delicious in the light to middle roasts, panela and molasses sweetness, winey apple and plum hints, a cinnamon note, tannic black tea and cranberry-like acidic impression. City to City+.	Wet Washed
Rwanda Dry Process Nyakabingo	Middle roasts move beyond molasses sweetness, to fruit and spice flavors, notes of berry-infused dark chocolate, plum, [illegible], and a hint of heart of palm in the finish. City+ to Full City.	Dry Natural

RECOMMENDER SYSTEM - RESULTS

- **Auto-Encoder**

Name	Description	Process
Kenya Embu Gakui Peaberry	Gakui's flavor profile is enlivened with fruit and spice notes like dried plum, date pieces, cinnamon stick, all spice powder, plum tea, tea-like tannic acidity, and some grapefruit bittering that...	Wet Washed
Colombia Caicedo Las Alegrias	A cup with intimations of dried fruit against a backdrop of rustic, unrefined sugar sweetness, hints of dried raisin and plum, tea-like tannic acidity. Chocolatey dark roast. City to Full City+.	Wet Washed
Rwanda Nyamasheke Gatare Peaberry	Aspects of semi-refined sugars, fruited acidity, laced with hints of warming spice, orange tea, and dried apple. Deep chocolate roast flavors with darker roast develoment. City to Full City+. Good...	Wet Washed

- **Word2Vec**

Name	Description	Process
Kenya Embu Gakui Peaberry	Gakui's flavor profile is enlivened with fruit and spice notes like dried plum, date pieces, cinnamon stick, all spice powder, plum tea, tea-like tannic acidity, and some grapefruit bittering that...	Wet Washed
Colombia Caicedo Las Alegrias	A cup with intimations of dried fruit against a backdrop of rustic, unrefined sugar sweetness, hints of dried raisin and plum, tea-like tannic acidity. Chocolatey dark roast. City to Full City+.	Wet Washed
Kenya Nyeri Kiruga AB	Depth of sweetness (scoring 9.5!), raw sugars, fruit jam hints, fig, dried berry and a spiced grape juice note as it cools. Moderate brightness and capable of berry-laden cocoa when roasted dark. ...	Wet Washed

RECOMMENDER SYSTEM - RESULTS

- **BERT:** [tanzania, molasses, plum, friar, washed]

Name	Description	Process
Burundi Commune Mutambu	Such a versatile Burundi, a neutral sweetness is accented by complex baking spices, creamed honey, loose leaf black tea and bittering cocoa when roasted dark. City to Full City+. Good for espresso.	Wet Washed
Guatemala Proyecto Xinabajul Donaldo Villatoro	An aromatic Guatemalan coffee with brisk acidity, toasted sugar sweetness, and flavor notes of warming spices, Earl Grey tea, dried plum and milk chocolate. City to Full City.	Wet Washed
Colombia Urrao Valle de Penderisco	Molasses, demurara sugar, moderate brightness, accents of berry and hibiscus flower tea. Dark roasts boast heavy-handed cocoa roast flavors and plum. City+ to Full City+. Good for espresso.	Wet Washed

- **BERT:** [~~tanzania~~, molasses, plum, friar, washed]

Name	Description	Process
Colombia Urrao Valle de Penderisco	Molasses, demurara sugar, moderate brightness, accents of berry and hibiscus flower tea. Dark roasts boast heavy-handed cocoa roast flavors and plum. City+ to Full City+. Good for espresso.	Wet Washed
Cameroon Caplami Java Cultivar	An interesting, thick bodied cup, notes of caramel and molasses, cinnamon, pipe tobacco, powdered orange drink, malted grains, aromatic woody dimension. City++ to Full City+.	Wet Washed
Burundi Rwiri Yagikawa	A medium-bodied coffee with silky mouthfeel, turbinado sweetness, opening up to hints of chamomile and roasted barley teas, clove powder, and a subtle whiff of orange in the nose. City to Full City.	Wet Washed

AREAS FOR IMPROVEMENT

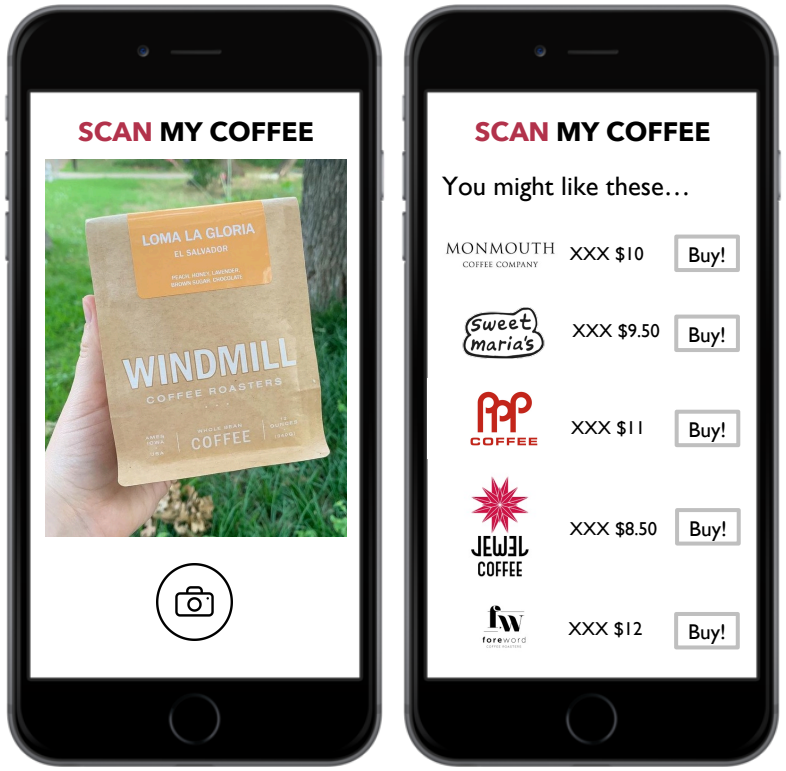
OCR

Infinite label design types	Further refinement needed to capture more types (e.g. handwriting, upside-down text)
Recognising > one word	'Friar plums' and not 'friar' and 'plums'
Unstructured Text	Name of coffee roasters captured (irrelevant for now)

Recommender System

Understanding context better	TF-IDF might be fit for purpose here, but it will not differentiate 'honey' (taste note) vs 'honey' (processing method)
Prioritisation by categories	'Decaffeinated, caturra, peach' – non-decaf options will show up. Must be decaffeinated > caturra/peach (hierarchical solution)

SUMMARY



- **Importance for text detection to work as well as text recognition**
- **TF-IDF provides a very fast keyword search solution with no need to train sophisticated models**

Future Developments

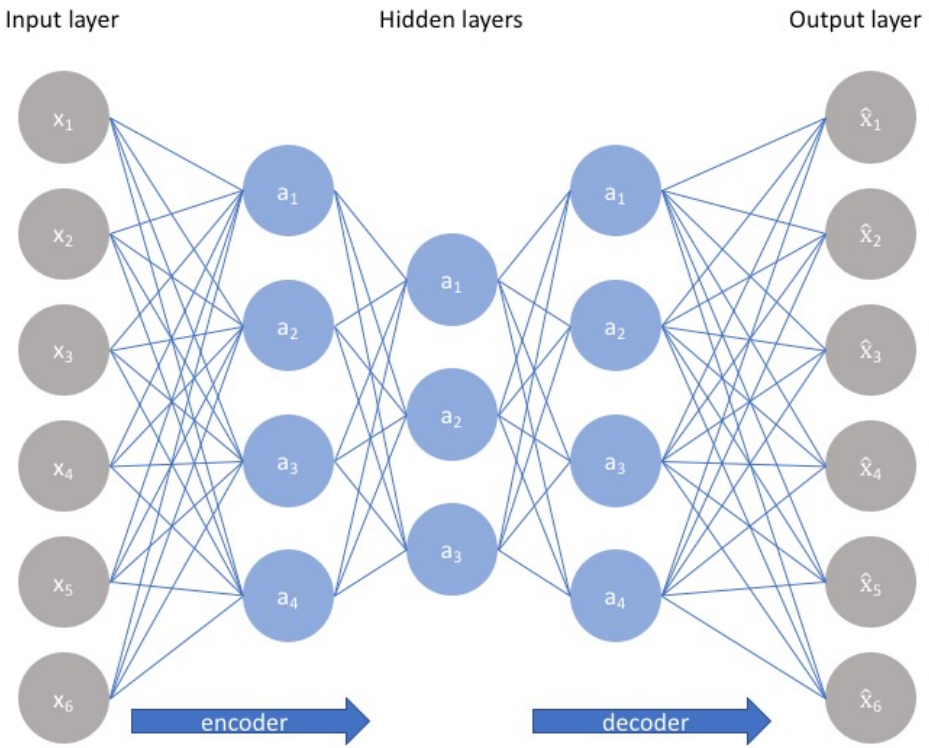
- **Scalable app solution – potential commercial opportunity for multiple retailers to come onboard**
- **Semantic models like BERT are more sustainable in the long term – e.g. manual queries, recommendations from user reviews**
  - **Train with larger dataset**
  - **TF-IDF + BERT?**

APPENDIX

FOR REFERENCE ONLY

PART II: RECOMMENDER SYSTEM

- **TF-IDF** (Baseline)
  - Fastest way to analyse relevance of word
  - Does not understand semantics or nuances
- **Auto-Encoder**
  - Unsupervised neural network
  - **Encode → Bottleneck → Decode**
  - Trained a simple 6 layer feed-forward model
  - Minimise reconstruction loss
  - Retain good information (similarities/differences)
  - Context-sensitive representation of the data



PART II: RECOMMENDER SYSTEM

- **Word2Vec**
  - Embedding technique that learns semantics of words
  - Trains against words that are neighbours (e.g. apple acidity) – windows
  - Key disadvantage: will not produce vectors for unseen data

- **BERT**
  - State-of-the-art NLP method, great with semantics
  - Bi-directional transformer
  - Attention mechanism – encode context from preceding and succeeding words
  - Does not treat words as independent input (vs TF-IDF, Auto-Encoder)
  - Deployed SentenceTransformers' *paraphrase-MiniLM-L6-v2* model