Keyword Spotting in Historical Documents — Report

Course: Pattern Recognition
Dataset: George Washington Collection

1. Preprocessing

1.1 Binarization

- Original grayscale document images were converted to **RGBA** and masked using the polygon data in SVG files.
- Word regions were extracted and isolated with **transparent backgrounds**, and then overlaid on white to simulate binarized words.

1.2 Word Cropping

- Each word was localized using its associated SVG path (`word images`) in the ground-truth location files.
- Coordinates were used to define bounding boxes and extract individual **word images** saved as `.png` files.

---

2. Feature Extraction

- Used a custom feature extractor (`extract_features`) to compute feature vectors for each word image.
- Features were stored for both training and validation words:
    - **Train words**: 2,433
    - **Validation words**: 1,293
    - **Total extracted features**: 3,726

---

3. Keyword Spotting via DTW

- One **query image** per keyword was selected from the training set.
- Each query was matched against **all validation word images** using **Dynamic Time Warping (DTW)**.
- DTW distances were computed between feature vectors, and ranked matches were generated.

---

4. Evaluation

4.1 Keyword Set

- `keywords.tsv` provided 35 target keywords.
- Only keywords present in the training transcription data were kept:
    - **Valid keywords**: 35
    - Matching based on normalized transcription text.

4.2 Metrics

- Used **Average Precision (AP)** per keyword.
- Final **Mean Average Precision (mAP)** over all 35 keywords:
    - **mAP = 0.3148**

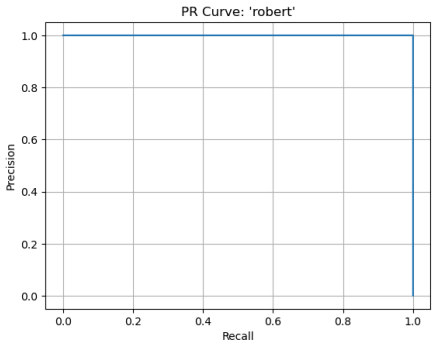
Top-5 Keywords by AP:

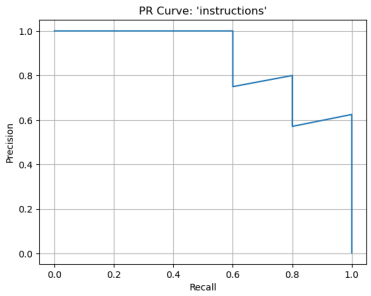
Keyword	AP

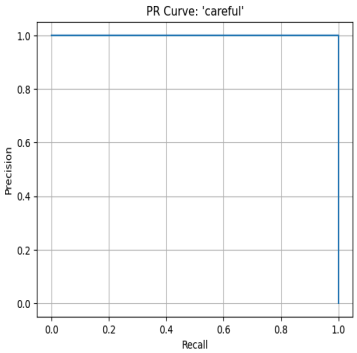
careful	1.0000
robert	1.0000
instructions	0.8850
major	0.7500
orders	0.6852

Precision–Recall Curves

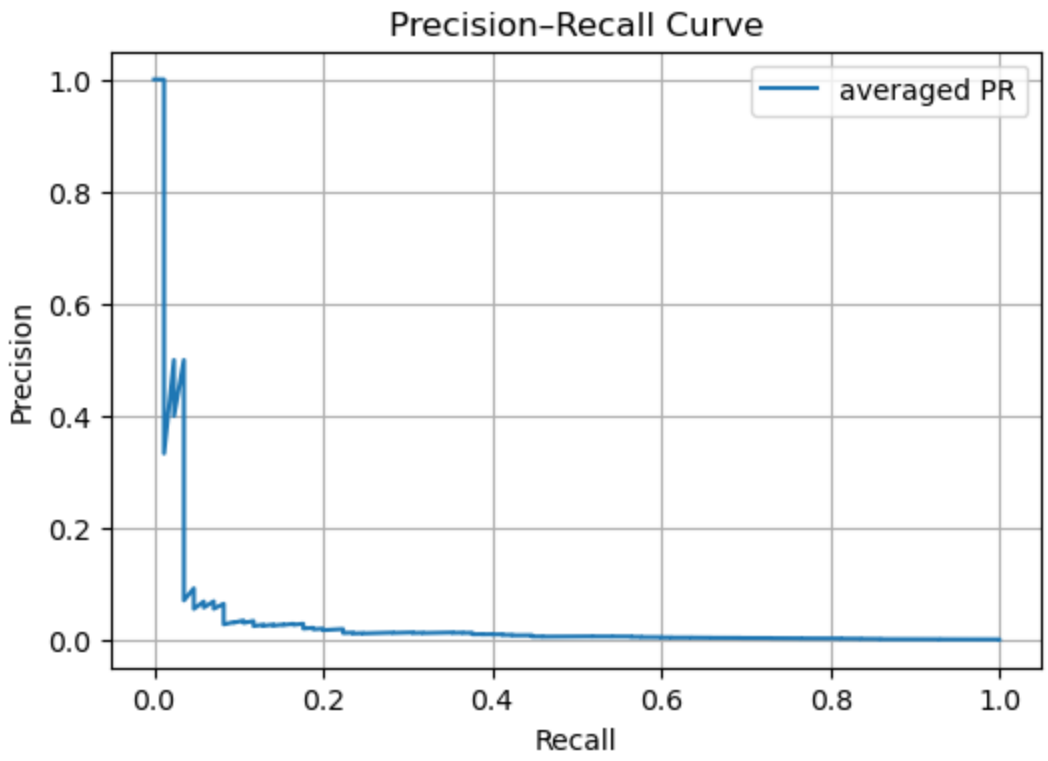
- Individual PR curves were plotted for:
    - **Top-3 keywords**: *careful*, *robert*, *instructions*







- An **aggregate PR curve** was generated by concatenating all results across all keywords.



---

5. Test Set Submission

5.1 Setup

- `keywordstest.tsv` provided both the keyword text and the **image ID** to be used as query.
- Test set: documents 305–309.
- DTW dissimilarities were computed between the query image and **all test word images**.

5.2 Output

Results saved in `test_output.tsv` in the required format:

```
keyword1 wordID1 dist1 wordID2 dist2 ...

keyword2 ...
```

- All test word distances were included and sorted by similarity (least to greatest).

  And `test_output_filtered.tsv` has top 8 word images for each keyword.