Dataset Preparation

We gathered only webpages in English. We were aiming at an average of 100 pages/genre – more of the pages belonging to genres we expected to be more difficult to identify and less of others. We ended up with at least 60 pages of each genre and nearly 200 pages belonging to the more common genres. Each page could be classified into multiple genres: either because all of it belonged to multiple genres or because it has sections belonging to different genres (the line between the two reasons was often blurred). Dataset preparation consisted of four steps:

1. It is desirable to be able to classify popular pages, pages that people actually search for. Therefore we input the most popular query from each of the five categories according to 2004 Year-End Google Zeitgeist into Google and used hits 31–60 were. The first 30 hits were skipped as suggested by Lim et al. (2004) to avoid too many commercial/promotional pages.

britney spears
orlando bloom
the simpsons
wallpaper
ebay

This gave us 150 pages.

2. In order to increase the diversity while retaining the popularity criterion, we input the most popular query of each weekly 2004 Google Zeitgeist into Google and used hits 31–35 (or 31–40 for pages topping the weekly Zeitgeist twice).

anna benson
anna kournikova
apprentice
ashlee simpson
cameron diaz
carly patterson
dimebag darrell
earth day
easter
election results
fahrenheit 911
golden globes
hellboy
howard stern
hurricane frances (2)
hurricane ivan
hurricane jeanne
iraq
janet jackson
jennifer aniston
jibjab
john kerry (2)
kentucky derby
kobe bryant
lance armstrong
lunar eclipse
madonna
mardi gras
martha stewart (2)
mega millions
memorial day
mother's day
mount st helens
nasa
nick berg

olympics
paul johnson
ray charles
ronald reagan
ryder cup
shrek 2
spiderman 2
tsunami
valentine cards
valentines day
yankees

This gave us 245 pages for a total of 395 pages.

3. We used 300 random pages from Mangle for a total of 695 pages.

4. Finally, we gathered pages belonging to genres insufficiently represented up to this point. This was mostly done by inputting genre-related queries into Google and using the relevant hits.

The webpages were gathered by a senior student of journalism. She was supervised by the authors and all the pages she considered difficult were discussed. Afterwards, the pages were classified again by another student, who was given only minimal supervision. All the pages on which the annotators disagreed were checked by the authors, who also decided the final classification.

Notes on the Gathered Data

Each item in the dataset has five fields:

- Filename – filename of the cached page.
- URL – URL of the original page.
- Source – one of the four steps of the gathering process.
- PrimaryGenre – genres assigned by the first student.
- SecondaryGenre – genres assigned by the second student.
- FinalGenre – the genre assigned by the authors, considered the correct genre.