Plan Overview

A Data Management Plan created using DMPTool

Title: Overbaked and Underproofed: An interactive exploration of the judging language in the Great British Baking Show

Creator: R C

Affiliation: City University of New York (cuny.edu)

Principal Investigator: Nuraly Soltonbekov

Data Manager: Teddy Manning

Project Administrator: Maria Baker

Contributor: RC

Funder: City University of New York (cuny.edu)

Template: Digital Curation Centre

Project abstract:

Overbaked and Underproofed: An interactive exploration of the judging language in the Great British Baking Show (GBBS), seeks to create and provide a platform that fosters a critical engagement with an aspect of language as used in a popular reality TV phenomenon.

By extracting and analyzing the vocabulary and language usage in the judging segments of two seasons of GBBS, the project seeks to probe the perceived paucity of evaluative language and then look at what the results might reveal about our culture's easy relationship with quick judgment. A comparison corpus from an American cooking show will further help clarify the particularities of evaluative language in the context of food judging competitions.

Via a website for engaged fans of GBBS (primarily) as well as interested linguists and academics, the project features visualizations of the analyzed judging language corpora, academic discourse around

the method and findings of the project, as well as an interactive Judge-this-bake Bingo game (populated with the most frequently used judgment expressions) that can be played while re/watching episodes of GBBS. Overbaked & Underproofed aims to playfully induce a shift in popular media consumption by bringing consciousness to the use of evaluative expressions and the often unarticulated frameworks of judgment, ultimately producing a new and expanded literacy.

Start date: 02-19-2023

End date: 05-30-2023

Last modified: 02-28-2023

Overbaked and Underproofed: An interactive exploration of the judging language in the Great British Baking Show

Data Collection

What data will you collect or create?

A corpus dataset of transcripts from The Great British Baking Show (Season 2 and Season 12) will be created. They will be in .txt file format. We are expecting one file per episode per season. There are 10 episodes per season, so we are expecting 20 files. Each file should be around 50K in size.

How will the data be collected or created?

Data will be collected by locating transcripts of relevant TV shows online, then cleaning the data of names &/or non-alphanumeric characters, etc. We will store the files in folders grouped by season. We will name the files by their season number, followed by their episode number. We will conduct a high-level manual review of the text file to ensure they are complete.

Documentation and Metadata

What documentation and metadata will accompany the data?

Ethics and Legal Compliance

How will you manage any ethical issues?

We are not anticipating any ethical issues related to this project at the moment as this dataset is publicly available information from a TV show. There will also be no personal identifiable information (PII) in this dataset that we will have to handle as names included in this dataset are public personalities that have appeared in a public TV show.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

Copyright shouldn't be an issue as this project counts as transformative, ie memetic, however, a takedown notice should be included on our project page.

Storage and Backup

How will the data be stored and backed up during the research?

Data will be stored on an organized Google Drive as well as a GitHub repository.

How will you manage access and security?

Security is a non-issue for this project as it does not include sensitive data.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The data will be retained and shared for three years until project sunset.

What is the long-term preservation plan for the dataset?

The data will be retained and shared for three years until project sunset. The project manager will be responsible for keeping the data alive.

Data Sharing

How will you share the data?

We will make the data shareable via Github.

Are any restrictions on data sharing required?

There is no restriction on data sharing required.

Responsibilities and Resources

Who will be responsible for data management?

Our Corpus Data Manager will be responsible for data management.

What resources will you require to deliver your plan?

A team of project manager, web developer, NLP analyst, and researcher is required to deliver our plan.

Planned Research Outputs

Text - "Text analysis"

Data paper - "NLP linguistic paper"

Interactive resource - "Website"

Image - "Bingo Game"

Dataset - "Text Data"

Software - "NLP Code"

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Text analysis	Text	2023-04-30	Open	None specified	10 MB	None specified	None specified	No	No
NLP linguistic paper	Data paper	2023-04-30	Open	None specified	1 MB	None specified	None specified	No	No
Website	Interactive resource	2023-04-30	Open	None specified		None specified	None specified	No	No
Bingo Game	Image	2023-04-30	Open	None specified	1 MB	None specified	None specified	No	No
Text Data	Dataset	2023-04-30	Open	None specified		None specified	None specified	No	No
NLP Code	Software	2023-04-30	Open	None specified		None specified	None specified	No	No