Using NLP and machine learning to classify cuisine type of recipes based on their ingredients.

Sunella Fernando | 21052642 | UFCFMJ-15-M MLPA Coursework Assignment

**Abstract;** This study attempts to identify the type of cuisine of a recipe based on its ingredients. A recipe dataset from Kaggle, originally scraped from the aggregation website Yummly was used for training and testing (Kaggle, 2016). The scope was limited to the ingredient name; quantity and preparation methods were not considered.

Source code; <link>

Keywords; NLP, machine learning, food, cuisine, classification.

# Introduction

Food is a basic human need. The accessibility of a variety of recipes on the internet from multiple different locations and cultures transcends barriers and opens up avenues to explore the world for those without the means and ability to travel; it helps bring people together by way of shared experience.

Submitting recipes to most popular online sources require filling in multiple fields manually. This can be discouraging as it might be difficult for amateur cooks submitting recipes to identify which type of cuisine their recipe is likely to align with, and even experienced authors might be put off by the number of input fields to fill in (Krug, 2014). Deciding what type of recipe to search for or what to cook considering the ingredients that are available is a familiar problem to many, as evidenced by the ever-increasing Google trends graph for relevant search terms (Google Trends, 2004-2022). Low quality recipes that are uploaded to the public domain without proper moderation or tagged with incorrect labels and cuisine types in an attempt to get more ‘views’ reduces the quality of recipe aggregation sites, misleads the target audience and might even contribute to increase food waste.

Developing a way to automatically predict cuisine type could help mitigate some of these issues, by potentially simplifying recipe submission processes, helping decide what type of cuisine to prepare with ingredients at hand, identifying low quality ‘clickbait’ recipes, and ultimately improving the quality and variety of food experiences available to all.

# EDA and Pre-processing

Some exploratory analysis was performed to better understand the dataset and identify what pre-processing steps would be required.

The Kaggle ‘What’s Cooking?’ dataset (Kaggle, 2016) contains two separate json files for training (tagged) and testing (untagged). The tagged training dataset which consists of 39774 recipes from 20 types of cuisines was used in this work. The dataset does not include any ‘fusion’ recipes (i.e. each recipe belongs to a single distinct cuisine type), which simplifies the problem, even though this may differ in real-world contexts.

The number of recipes belonging to each cuisine type is not equally distributed; Italian is most prevalent with 7838 instances, followed by Mexican (6438 recipes), with Russian and Brazilian cuisine being on the lower end with 489 and 467 recipes respectively (Figure 1).

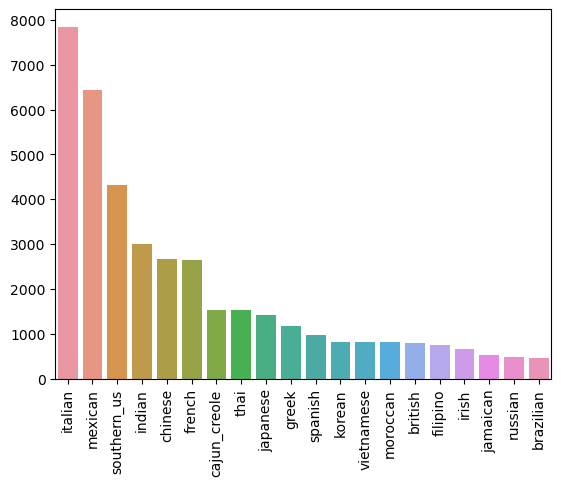


Figure 1- Cuisine counts

There are 6714 unique ingredients with the most common being salt (18049 occurrences), onions(7972), olive oil(7972), and water(7457), and least common being more specialized ingredients with brand associations, specific cuts of meat or rare vegetables that occur only once such as lop chong, Kraft cheese crackers, Oscar meyer cotto salami, Lipton iced tea, and tongue, which each occur just once. (Figure 2).

Ingredient word cloud.


Figure 2- Ingredient word-cloud

A quick analysis was done to evaluate whether the number of ingredients could be used as a feature in itself (i.e. whether there were any discernible patterns where, for example recipes of a certain type of cuisine were likely to have longer/shorter lists of ingredients than others), but the average number of ingredients appears to be similar for almost all cuisine types, thus making the ingredient count an unlikely candidate for a ml model to learn from.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cuisine** | **Average** | **Min** | **Max** | **SD** |
| greek | 10.18 | 1 | 27 | 3.72 |
| southern\_us | 9.63 | 1 | 40 | 3.860 |
| filipino | 10 | 2 | 38 | 3.85 |
| indian | 12.71 | 1 | 49 | 5.01 |
| jamaican | 12.21 | 2 | 35 | 4.76 |
| spanish | 10.42 | 1 | 35 | 4.16 |
| italian | 9.91 | 1 | 65 | 3.80 |
| mexican | 10.88 | 1 | 52 | 4.65 |
| chinese | 11.98 | 2 | 38 | 4.04 |
| british | 9.71 | 2 | 30 | 4.16 |
| thai | 12.55 | 1 | 40 | 4.41 |
| vietnamese | 12.68 | 1 | 31 | 5.25 |
| cajun\_creole | 12.62 | 2 | 31 | 4.61 |
| brazilian | 9.52 | 2 | 59 | 5.55 |
| french | 9.82 | 1 | 31 | 4.14 |
| japanese | 9.74 | 1 | 34 | 4.24 |
| irish | 9.3 | 2 | 27 | 3.70 |
| korean | 11.28 | 2 | 29 | 3.87 |
| moroccan | 12.91 | 2 | 31 | 4.79 |
| russian | 10.22 | 2 | 25 | 4.05 |

Table 1-Ingredient length distribution

It was observed that multiple cuisine types had occurrences of recipes with just one or two ingredients which might need to be removed. However on further analysis, it was decided to retain such

# Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file “MSW\_USltr\_format”.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive.”
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter,” not “webers/m2.” Spell units when they appear in text: “...a few henries,” not “...a few H.”
* Use a zero before decimal points: “0.25,” not “.25.” Use “cm3,” not “cc.” (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

*a**b*    

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1),” not “Eq. (1)” or “equation (1),” except at the beginning of a sentence: “Equation (1) is ...”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o.”
* In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset,” not an “insert.” The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively.”
* In your paper title, if the words “that uses” can accurately replace the word using, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect,” “complement” and “compliment,” “discreet” and “discrete,” “principal” and “principle.”
* Do not confuse “imply” and “infer.”
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”
* The abbreviation “i.e.” means “that is,” and the abbreviation “e.g.” means “for example.”

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is “Heading 5.” Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract,” will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1,” “Heading 2,” “Heading 3,” and “Heading 4” are prescribed.

## Figures and Tables

### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1,” even at the beginning of a sentence.

1. Table Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. *(Table footnote)*
2. Example of a figure caption. *(figure caption)*

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization,” or “Magnetization, M,” not just “M.” If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization (A ( m(1),” not just “A/m.” Do not label axes with a ratio of quantities and units. For example, write “Temperature (K),” not “Temperature/K.”

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi resolution TIFF or EPS file with all fonts embedded) because this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

1. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
2. I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
3. K. Elissa, “Title of paper if known,” unpublished.
4. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
5. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
6. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.