Project Background

The User Intent classifier that I am developing for my final project began at my company (Ask.com) last year as an open-ended exploration of our keyword portfolio. Previously, the keywords in this portfolio had been examined for the purpose of building thematic classifiers that recognized keywords were about specific topics (such as “Health” or “Sports), and had also been segmented in various ways based on user interaction in order to assess their value. This “deep dive” (as it was termed) was initiated to gain a deeper knowledge of the character and behavior of keywords in the portfolio.

Based on my initial analysis of the portfolio keywords, I developed a set of User Intent classes, helped select a segment of the keyword portfolio for labeling, and created labeling guidelines for an outside vendor, Crowdsource. Using my guidelines, Crowdsource performed a larger-scale labeling task of ~100k Health-related keywords for User Intent. This data set is used in my final project in an attempt to create a user intent classification model.

Initial Exploratory Analysis

For my initial analysis, I looked at segments of the keyword portfolio, including:

- Top Performing Keywords (Based on clicks and traffic)
- Samples of our top performing thematic categories: Health, Home&Garden, Vehicles, Travel
- Random Cross-Categorical Samples

These samples were chosen to represent the overall breadth of the portfolio, but we also wanted to give special consideration to the keywords that users engaged with most frequently (and therefore the keywords with the greatest impact on our company).

I used regexes and other techniques uncover common patterns in these keywords. Based on these patterns, I proposed a set of User Intent classes for the keywords that would classify the keywords from the user’s perspective. The classes would attempt to answer the question: “What does the user want to obtain or hope to accomplish when typing in this keyword?” The details of these User Intent classes are included in the “Data Set Description” section below.

Data Set Selection

Based on conversations with product stakeholders and some experimental labeling over various thematic categories of the keyword portfolio, we decided to use a sample of keywords that were classified as Health by our in-house thematic classification system. The Health category has a strong definition and is thematically consistent-- it is almost certain that all keywords categorized as Health by the thematic classification system are related to health and wellness. These keywords are also mostly unambiguous-- we would be less likely to run into labeling errors that might occur in other categories. It also seemed logical to focus our efforts to understand user intent on a category with high volume and that contains keywords that exhibit a high degree of user engagement.

Data Set Description

The data used for this User Intent classifier consists of 90,835 keywords from the keyword portfolio that were classified as Health by our in-house thematic classifier. These keywords were then labeled for User Intent by Crowdsource, based on the classes described below:

- **Navigational**
  - User wants to navigate to a website, or information hosted on a specific website. Business names and government agencies without additional context are included in navigational.
  - Example: aetna login, myhealth com
- **Resource**
  - User wants to find a specific type of resource or media type. The resource can be confined to the web, such as reviews, calculators, or look ups, OR the resource can be something that can be printed out, downloaded, or viewed.
  - Example: diabetes blood chart, pictures of skin rash
- **Generic**
  - The keyword has a clear topic, but no additional context which would indicate what the users would like to know or see. Generic seems to imply the user request: "Show/Tell me *anything* about Topic"
  - Example: breast cancer, facts about cellulite
- **Direct Answer**
  - The keyword indicates that the user has a specific question, and is looking for a specific answer. The length of the answer can vary from a single word to an entire article. This label includes requests for lists, or keywords that would prompt a list-like response. Direct Answer implies the user request: "Show/Tell me *X* about Topic". Example: list of flu symptoms, what is the difference between diabetes I and II
- **Guides & Instructions**
  - User's intent is to find the steps to accomplish a task or project. This includes instructions, guides, and recipes.
  - Example: how to treat wasp sting, change wound dressing
- **Transactional (Shopping Intent, or Product Name)**
  - This category can be applied to keywords with any explicit transactional component--keywords about buying, selling, renting, prices, costs, values, etc. It also is applied to all product names, any item that can be bought or sold. For this specific use case, it made sense to include product names along with explicit shopping requests.
  - Example: truvada cost, buy viagra

The data also includes another labeled class called **Authority.** For this label, we asked the Crowdsource evaluator to use their personal judgment to select the level of expertise they would want or expect from a responder to a given keyword.

Data Set Preprocessing

I am very fortunate that although a great deal of work went into initial analysis and project definition, the data set itself is remarkably clean.

	CS_ID	Keyword_ID	Keyword	Intent_1	Authority
0	CS_0001	3486	poison oak pictures	Resource	General
1	CS_0002	3486	best foods for hypothyroidism	Direct Answer	Expert
2	CS_0003	3486	kidney stones in women	Generic	Expert
3	CS_0004	3484	what spider bites look like	Resource	General

I dropped columns CS_ID, Keyword_ID, and Authority, as these columns will not be used in the classification model. In the Intent_1 column, I converted the User Intent labels to integer values.

For the keywords, I used a function that tidied them up by lowercasing them (they are already lowercased, I think, but why not be thorough?) and stripping out non-letters. At this step, I also implemented NLTK's built-in stopword list to pull out English function words that occur frequently in documents, but often do not contribute meaning (i.e. articles "a", "the").

After using this function, I found out that Sci-Kit's CountVectorizer handles a lot of text pre-processing and in all likelihood this little function is pointless. Oh well.

In order for my keywords to be used in a model, they needed to be transformed into feature vectors. I used Sci-Kit's CountVectorizer to convert the keywords into a matrix of feature counts. Although token count can be really useful, it's not exactly the best indicator that a given token is important--some words might just appear more commonly than others. Thus, I implemented Sci-Kit's TfidfTransformer. Tf-Idf stands for "Term Frequency times Inverse Document Frequency". With Tf-Idf, the importance of a token increases proportionally to the number of times a word appears in the document ("term frequency"), but is offset by the frequency of the word in the corpus "inverse document frequency".

Building a Model

Before building my model, I had a conversation with Justin during office hours and he suggested that I first build a binary model that would tackle the most prominent User Intent classes in my data--Generic and Direct Answer. After building that model, I would pipe the output to a second model that would handle the smaller, less representative intents: Resource, Guide, Navigational, and Transactional.

For this first classifier, I chose to attempt a Naive Bayes model. This first pass has an accuracy rate of 76%.

Next Steps

Understand NLP Methods and Approaches: I'm strong on analysis and know a lot about language (I'm a linguist), but I'm also really new to programming. For me, everything takes a lot of time but I needed to have some kind of draft. My primary goal for the next couple of weeks is to dive really deeply into Natural Language Processing, and understand the various approaches and tools available to me.

Data Exploration: Again, everything takes me a really long time, so this pass was mostly about making something go, not digging deeply into the language data itself.

- Are there types of terms or patterns that are prevalent in a given User Intent class?
- What kind of impact is NLTK's stopword list having on the performance of my model? The User Intent classifier isn't just classifying keywords into content topics, it should be getting clues about intent from the structure of keywords as well (so maybe some of those functional words are actually important).

Error analysis on the current model

Attempt other models besides Naive Bayes: Why did I pick Naive Bayes? Because it seemed like a solid approach for text classification, the research I did also indicated it might be a good choice (this or SVM, which we haven't learned about yet). I'd like to see how other model types would interact with this data, especially since the first classifier is a binary classifier.

Build the other classifier to tackle the smaller User Intents! Yay!