Fine-grained sentiment analysis

Problem Statement:

Given a set of labeled aspect specific sentiment expressions across variety of review sentences, find new sentiment expression spans on unseen sentences of the same aspect.

Task 1: Given a head aspect (HA), detect whether a sentence containing the HA **mentions an issue.**

Task 2: Given a HA and a sentence mentioning an issue, extract the **issue phrase boundary**.

Dataset

Domain and Head aspect.

earphone	gps	keyboard	mouse	mp3_player	router
cord	direction	keys	battery	button	connection
jack	screen	pad	button	interface	firmware
wire	software	range	pointer	jack	signal
	voice	spacebar	wheel	screen	wireless

Task I: Ideas tried and results

1. Processed text to fetch Issue and Non Issue Labels.
2. Removed stop words to improve accuracy.
3. Removed <i> and POS tags at the end of sentence.

Task I: Ideas tried and results

Example:

Domain: Earphone. Head aspect: wire.

I do however have some pain points: First, the "stethoscope" effect is actually quite bad unless you run the headset under your shirt and/or tighten the <i> wire under your chin <i> . [do, have, is, run, tighten] [bad]

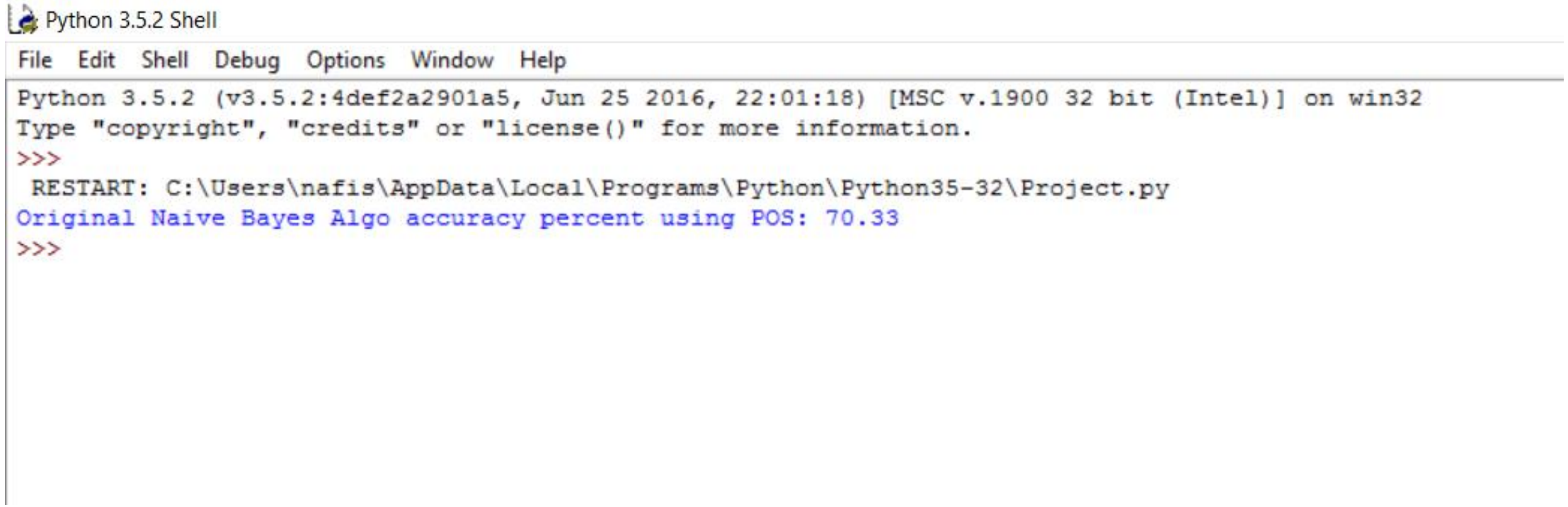
Make no mistake about it, the J3's are beautiful, solid aluminum construction, awsome clear coat wire that does not tangle. [are, does, tangle] [beautiful, solid, awsome, clear]

Sentence	Label
I do however have som…….<i> wire under your chin <i> . [do, have, is, run, tighten] [bad]	Issue
Make no mistake about …….	No Issue

Task I: Ideas tried and results

4. Implemented Features with Naïve Bayes Model
   - High frequency POS Tags

     Accuracy: 70%+



Task I: Ideas for future

1. Use some classification models from below and use a voter system
   - Multinomial NB
   - Bernoulli NB
   - Logistic Regression
   - Stochastic Gradient Decent Classifier.
2. Understanding the impact and setting user defined parameters in above models for optimum results. Example: Multinomial Parameters

Parameters:

alpha : float, optional (default=1.0)

Additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing).

fit_prior : boolean, optional (default=True)

Whether to learn class prior probabilities or not. If false, a uniform prior will be used.

class_prior : array-like, size (n_classes,), optional (default=None)

Prior probabilities of the classes. If specified the priors are not adjusted according to the data.

Task I: Ideas for future

3. Voter System:

Makes use of classification given by above mentioned classifiers to give confidence of final classification made.

Pending Implementation

1. 5 Fold cross validation
2. Other Features to be added(Phrase Chunk, Prefix, Suffix, Sentiment Polarity)

Pivot Features: We consider five feature families which take on a set of values:
POS Tags (T): *DT*, *IN*, *JJ*, *MD*, *NN*, *RB*, *VB*, etc.
Phrase Chunk Tags (C): *ADJP*, *ADVP*, *NP*, *PP*, *VP*, etc.
Prefixes (P): *anti*, *in*, *mis*, *non*, *pre*, *sub*, *un*, etc.
Suffixes (S): *able*, *est*, *ful*, *ic*, *ing*, *ive*, *ness*, *ous*, etc.
Word Sentiment Polarity (W): *POS*, *NEG*, *NEU*

Task II: Phrase Extraction-Understanding

1. Unsupervised Heuristic Baseline (UHB)
   - Identify words between head aspect and a negative sentiment OR head aspect, a positive sentiment and a negator.
2. Hidden Markov Model
   - Let $\boldsymbol{x}$ = ($x1$, … xn) denote the sequence of words in a sentence.
   - Let each observation xi has a label yi ∈ Y where Y= {B,I,O}.
   - The extraction task is to find the best label sequence yˆ that describes an issue.
   - Here we find $f(yi{-}1,yi,x)$ for each word in sentences

Issues Encountered & Doubts

1. Feature Identification

Category	Feature Template	Example of feature appearing in a sentence
1st order features $X_{i+j}; -4 \leq j \leq 4$ $X \in \{T, C, P, S, W\}$	W_{i+j}	$W_{i-1} = NEG$; previous term of head aspect is of NEG polarity, … have this terrible *voice* on the…
	S_{i+j}	$S_{i-2} = ing$; suffix of 2nd previous term of head aspect is "ing", …kept dropping the *signal*…
	…	…
2nd order features $X_{i+j}, Y_{i+j}; -4 \leq j \leq 4$ $X, Y \in \{T, C, P, S, W\}$	$T_{i+j}, T_{i+j'}$	$T_{i-2} = JJ, T_{i-1} = VBZ$, …frequently drops *connection*…
	$T_{i+j}, C_{i+j'}$	$T_{i+2} = RB, C_{i+3} = ADJP$; …*screen* is too small
	…	…
3rd order features $X_{i+j}, Y_{i+j}, Z_{i+j}$; $-4 \leq j \leq 4$ X, Y, Z $\in \{T, C, P, S, W\}$	$T_{i+j}, S_{i+j'}, T_{i+j''}$	$T_{i+2} = JJ, S_{i+4} = un, T_{i+4} = JJ$; …*screen* is blank and unresponsive…
	…	…

Table 2: Pivot Feature Templates. The subscript i denotes the position of the issue subject (HA) which is italicized and the subscript j denotes the position relative to i.

Job responsibilities

Mohit	Nafisa
Understanding the impact and setting user defined parameters in above models for optimum results	5 Fold cross validation
Voter System	Other Features to be added(Phrase Chunk, Prefix, Suffix, Sentiment Polarity)
Hidden Markov Model	Unsupervised Heuristic Baseline (UHB)