For Part 1 :
1.

Entropy of outcome Y = 1

$$H(Y) = 1$$

HasJob

When HasJob = 0 w = 3/8
Y = 0 w/ P = 2/3
$\frac{2}{3} log_2(\frac{3}{2}) = .389975$

Y = 1 w/ P = 1/3
$\frac{1}{3} log_2(3) = .528321$ $.389957 + .528321 = .918$

When HasJob = 1 w = 5/8
Y = 0 w/ P = 3/5
$\frac{3}{5} log_2(\frac{5}{3}) = .442179$

Y = 1 w/ P = 2/5
$\frac{2}{5} log_2(\frac{5}{2}) = .52877$ $.442179 + .52877 = .971$

Info Gained for HasJob: $H(Y|X) = .918(\frac{3}{8}) + .971(\frac{5}{5}) = .95$

Info Gained: $H(Y) - H(Y|X) = 1 - .95 = 0.05$

HasFamily

When HasFamily = 0

w = 1/2
Y = 0 w/ P = 1/4
$\frac{1}{4} log_2(4) = .5$

Y = 1 w/ P = 3/4
$\frac{3}{4} log_2(\frac{4}{3}) = .31128$ $.5 + .311 = .81$

When HasFamily = 1

w = 1/2
Y = 0 w/ P = 3/4
$\frac{3}{4} log_2(\frac{4}{3}) = .81$

Y = 1 w/ P = 1/4
$\frac{1}{4} log_2(4) = .5$

$.311 + .5 = .81$

Info Gained for HasFamily: $H(Y|X) = .81(\frac{1}{2}) + .81(\frac{1}{2}) = .81$
Info Gained: $H(Y) - H(Y|X) = 1 - .81 = 0.19$

IsAbove30Years

When IsAbove30Years = 0

w = 1/4
Y = 0 w/ P = 1/2
$\frac{1}{2}log_2(2) = .5$

Y = 1 w/ P = 1/2
$\frac{1}{2}log_2(2) = .5$

$.5 + .5 = 1$

When IsAbove30Years = 1

w = 3/4
Y = 0 w/ P = 1/2
$\frac{1}{2}log_2(2) = .5$

Y = 1 w/ P = 1/2
$\frac{1}{2}log_2(2) = .5$

$.5 + .5 = 1$

Info Gained for IsAbove30Years: $H(Y|X) = 1(\frac{1}{4}) + .1(\frac{3}{4}) = 1$
Info Gained: $H(Y) - H(Y|X) = 1 - 1 = 0$

$\therefore$ HasFamily is gives the most information and is the most useful for the first split

For Part 2:

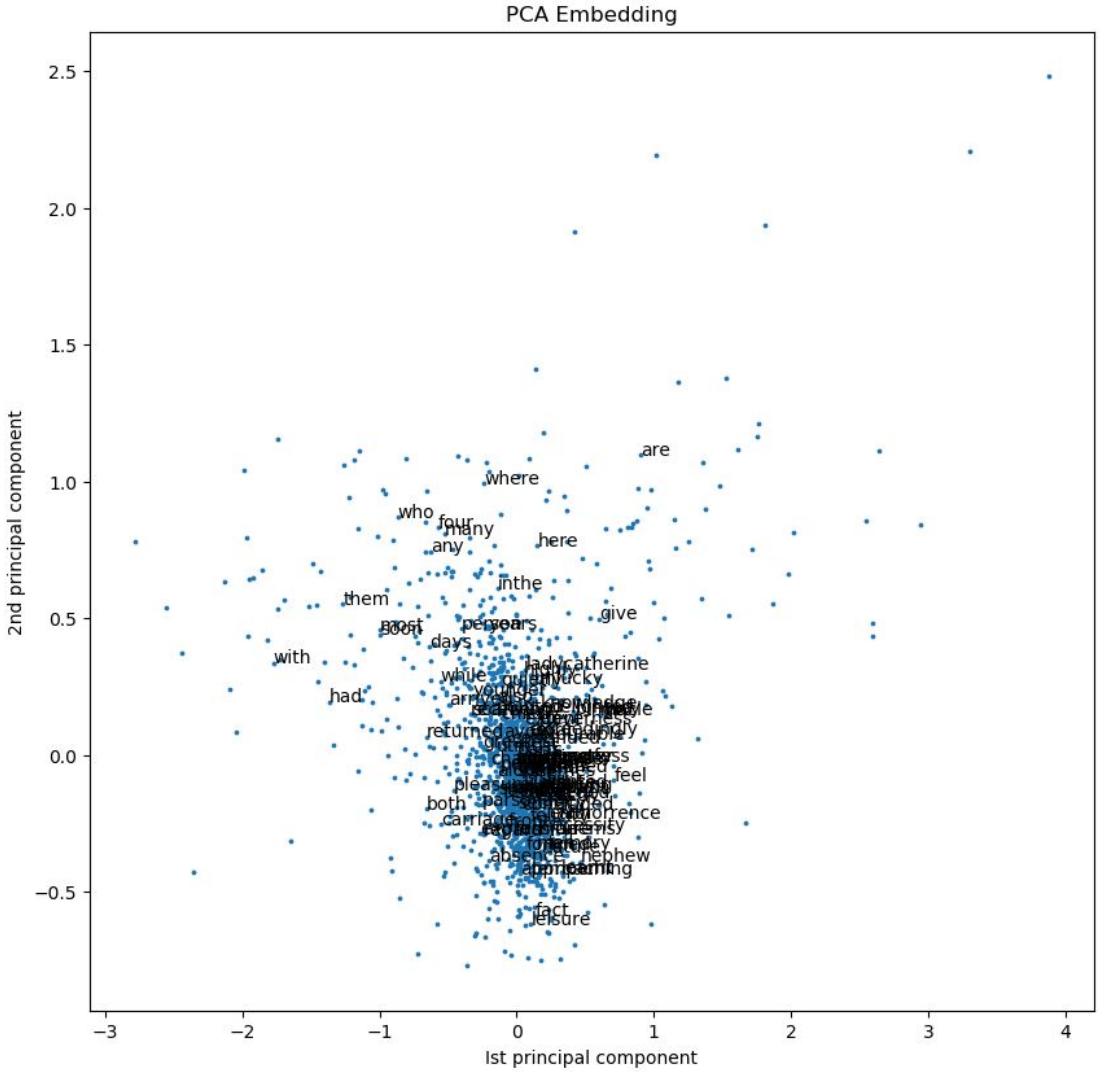
1.

1) Bag of Words and Word2Vec models refer to ways in which the contents of a document are classified. The Bag of Words model creates unordered list of words without syntactic, semantic and POS tagging. Word2Vec models creates features and weights assigned to those features and is algebraic in nature. Bag of Words does not take word context into consideration but is simpler. Word2Vec has a steeper learning curve.
2) A word vector is a vectorization of a word that transforms the word into elements with weights attached. A word embedding is a lookup table, that given a word, returns the its corresponding array. Word frequency, vector dimensionality and context are some of the factors influencing word embedding.
3) A corpus is a history of written text or spoken words which is used for analysis and validation.

2.

a. I first imported the sentences and cleaned the data. Then I used a Word2Vec model and PCA decomposition to get two principal components of the feature space.
b. Report the vocabulary count, embedding size, number of training iterations in your modelling.
Vocab length: 1941
Corpus count: 5370

c. Report and explain your observation from the visualization of word vectors.



d. The most similar word to "england" is bye with a similarity of 0.810457706451416
Similarity between elizabeth and man -0.018027484838473818
Similarity between elizabeth and girl 0.4327758314448439
Between the words "england", "queen" and "king", the word king does not match

For Part 3:
1. These can be found in the code.