**Storyboard for Module 1.4 – Working with corpus data**

**Last revised: 26th June 2019**

**Roles:**

* Curriculum design:
* Story writer:
* Jupyter notebook coder/resource finder:
* Tester/rewriter:

**Time scale:**

* Target date for completion of Modules 1.1 to 1.4 is end of July 2019

**Working environment:**

* MS Notebooks + Slack ?

**Module 1 theme: Basic Python concepts for corpus processing (3 + 1 weeks in total)**

Objectives of Module 1: (i) Understand the basic concepts of Python programming using Jupyter Notebooks in the Microsoft Azure environment (‘*Azure Notebooks*’). This will include variables, collection types, loops, control structures and functions. Along the way students will learn to (ii) Write efficient regular expressions to solve text-based extraction tasks such as sentence segmentation, part-of-speech tagging and building a simple ELIZA-like chatbot; (iii) Apply the edit distance algorithm to text sequence problems; (iv) Work with corpus data to calculate statistics using loops, dictionaries and counting; (v) Consolidate understanding of commonly used evaluation metrics such as accuracy, precision, recall and F-score.

**Module 1.4 (week 3) – Working with Corpus Data**

Number in brackets show estimated time to complete (needs testing). **Text in bold** highlights points of self-assessment or formal assessment. **Text in blue** corresponds to topics covered in the lectures.

Module 1.4 (2 hours total)

**Short pre-module quiz** (<-- activate knowledge of n-grams and probabilities from the lecture)

Counting and plotting

1. Counting vocabulary
2. Creating a function for the type-token ratio (<-- consolidation of the work on functions we did last week, compare type-token ratio on a corpus of Shakespeare)
3. Calculate and plot a frequency distribution (<-- introduces importing a module for graphs)
4. Counting bigrams
5. **Practice Quiz:** counting (5 multiple choice questions that the students can use for formative assessment)

N-grams and probabilities

1. Estimating bigram probability
2. Estimating a sentence probability (<--applying Chain Rule and Markov Assumption)
3. **Practice Quiz:** estimating probability (5 multiple choice questions that the students can use for formative assessment)
4. Generating text using a language model (<-- compare perplexity using a 2-gram, 3-gram and 4-gram model on a corpus of Shakespeare?)

**Quiz 3:** Working with corpus data assessment (4 questions to be handed in on Week 4)