PAPER CODE NO.	EXAMINER:	Prof S Maskell	
COMP 529	DEPARTMENT:	**CS**	TEL. NO: 44573

FIRST SEMESTER EXAMINATIONS 2017/18

BIG DATA ANALYTICS

TIME ALLOWED: Two Hours

INSTRUCTIONS TO CANDIDATES

All candidates should answer **all three** questions

The numbers in the right hand margin represent an **approximate guide** to the marks available for that question (or part of a question). Total marks available are 100.

Additional Information:

None

1) a) i) Which V is Hadoop primarily designed to cater for? **1**

ii) Where does a programmer articulate that Hadoop is running in standalone mode? **1**

iii) What are the names of the daemons that act as master and slave in a cluster running Hadoop? Be clear which are which. **4**

iv) What is the name of the daemon responsible for maintaining a backup of the information describing where each file is being stored? **1**

v) Which feature of Hadoop makes it necessary to use a portable programming language such as java? **1**

vii) What does HDFS stand for? **1**

viii) A HDFS system uses 64 bit addresses and each block is 64 Megabytes. What is the maximum volume size for this HDFS system? **4**

ix) In what ways does a FAT32 hard disk differ from such an HDFS system? **2**

x) What are the inputs to and outputs from a reducer within a MapReduce job? **4**

b) i) Which V is Storm primarily designed to cater for? **1**

ii) Where does a programmer articulate that Storm is being run in local mode? **1**

iii) What are the names of the daemons that act as master and slave in a cluster running Storm? Be clear which is which. **2**

iv) What is the name of the daemon responsible for ensuring the stability of a Storm cluster? **1**

Question 1 continues overleaf.

v) A topology comprises two spouts and three bolts. Assume one spout generates a stream of images and the other spout generates a stream of 30 millisecond audio chunks. Assume one bolt performs lip-reading, one performs speech recognition and the third bolt aligns two streams of text. Draw a diagram describing the topology. Label all spouts and bolts. Annotate all streams with the information being transmitted. **5**

vi) Name a middleware product that is not Storm and that is designed to support streaming analysis. **1**

c) Aside from Volume and Velocity, state the other two Vs of Big Data and explain what they each mean. **4**

Total

34

2. We wish to model how social media can be used to detect drugs' side effects. We assume that: T is true if someone takes a drug (and false otherwise); O is true if someone is old (and false otherwise); U is true if someone is a social media user (and false otherwise); S is true if a user experiences a side-effect (and false otherwise); D is true if a social media user is detected as experiencing a side-effect when taking a drug (and false otherwise).

a) i) How many numbers would be needed to store $P(T,S,U,D)$ in general? **1**

ii) Write $P(T|D)$ in terms of $P(D|T)$ and $P(D)$. **2**

iii) How many additions and multiplications are needed to calculate $P(T|D)$ in this general case when D is true and T is true? **4**

iv) How many further mathematical operations are needed to calculate $P(T|D)$ when D is true and T is false? **1**

b) A model is postulated for which $P(T,S,U,D)=P(T)P(U)P(S)P(D|S,T,U)$.

i) Draw a Bayesian Network to describe this postulated model. Clearly label all nodes. **4**

ii) How many numbers would be needed to store the probabilities parameterising the postulated model? **2**

iii) Write $P(D|T)$ in sum-product form for this model. **2**

iv) Write $P(T|D)$ as a ratio of two sum-products. **3**

v) How many additions and multiplications are needed to calculate $P(T|D)$ using this postulated model when D is true and T is true? **2**

Question 2 continues overleaf.

b) Another model assumes that P(T,O,S,U,D)=P(T)P(O,U)P(S)P(D|S,T,U).

i) Draw a Bayesian Network to describe this model. Clearly label all nodes. **4**

ii) Write an expression for P(T,O,S,U,D) in terms of P(O|U). **2**

iii) Describe (in words not equations) the operation of belief propagation in the context of this model. **3**

iv) Describe (in words not equations) the operation of Pearl's algorithm in the context of this model. **3**

Total 33

3. a) You are part of a team building a predictive text system for a MOOC that helps people to learn Swedish. The system has been configured to consider a dictionary of 600,000 unique words. A Hidden Markov Model (HMM) is to be used to process the incoming stream of text. The Levenshtein distance is to be used as a way of quantifying the (non-negative) number of changes (e.g., additions and deletions) to one string to transform it into another string. The likelihood is defined, using Levenshtein distance, such that the likelihood is high for words in the dictionary that are similar to the current text string and low for words that are dissimilar to the current text string. The strings that have been processed up to and including now are $y_{1:t}$ and the true current word is x_t.

i) How many numbers are used to parameterise are the transition matrix? **2**

ii) What two reasons would motivate the approximate the transition matrix as sparse? **2**

iii) What would be a disadvantage of such an approximation? **1**

iv) Write an equation that expresses the fact that the state is Markov. **2**

v) Write an equation that expresses the fact that the current state is a sufficient statistic of the past in terms of predicting the measurement. **2**

vi) Write an equation for $p(x_t|y_{1:t})$ in terms of $p(x_{t-1}|y_{1:t-1})$, $p(x_t|x_{t-1})$ and $p(y_t|x_t)$. Clearly label the posterior, the prior, the likelihood and the dynamic model. **8**

vii) Draw a Bayesian Network for the joint distribution of 10 true words and 10 observed text strings. Clearly label one arc that represents a likelihood and one that represents an instance of the dynamic model. **7**

viii) Why would the output from the current time-step not be the same as the most likely word given the most likely word from the previous time-step? **2**

Question 3 continues overleaf.

b) i) What properties of the dynamics and likelihood are such that the Kalman filter exactly characterises the current uncertainty in the state given some historic data? **2**

ii) What two parameters are passed between consecutive iterations of a Kalman filter? **2**

iii) What approximation is used by each of the Extended Kalman filter, Unscented Kalman filter and particle filter? **3**

Total

33