Questions & Answers on Chapter 2

1. What are the key stages of a generic pipeline for NLP system development? 🚀

1. **Data Acquisition** 📥
2. **Text Cleaning** 🧹
3. **Pre-Processing** ✂️
4. **Feature Engineering** 🛠️
5. **Modeling** 🧠
6. **Evaluation** 📊
7. **Deployment** 🚀
8. **Monitoring & Updating** 🔄

---

2. How can we get data required for training an NLP technique? 📂

- **Public Datasets** 🌍 (Kaggle, Google Dataset Search).
- **Scraping** 🕸️ (e.g., forums, news sites).
- **Product Intervention** 🤖 (design apps to collect user data).
- **Data Augmentation** 🔄:
  - Synonym replacement 📖 ("happy" → "joyful").
  - Back translation 🌐 (Translate EN → FR → EN).
  - Bigram flipping 🔄 ("machine learning" → "learning machine").
  - Adding noise ⌨️ ("hello" → "hlelo").

---

3. List the different data augmentation methods? 🎨

- **Basic**:
  - Synonym replacement 📖
  - Back translation 🌐
  - Bigram flipping 🔄
  - Replace entities 🏙️ (e.g., "London" → "Paris").
- **Advanced**:
  - **Snorkel** 🤿: Auto-label data with rules.
  - **EDA/NLPAug** 📚: Libraries for synthetic data.
  - **Active Learning** 🧠: Label only uncertain data.

---

4. Data can be collected from PDF files, HTML pages, and images, how this data can be cleaned based on their sources? 🧼

- **PDF**: Use `PyPDF2` to extract text.
- **HTML**: Remove tags with `BeautifulSoup`.
- **Images**: Extract text via OCR (e.g., Tesseract).

---

5. Using dot (.) to segment sentences can cause problems, explain how? ❗

- **Problem**: Splits abbreviations (e.g., "Mr. Smith" → "Mr" + ".").
- **Solution**: Use NLP libraries (spaCy, NLTK) to detect context.

---

6. What are the frequent steps in the data pre-processing phase? ✨

1. Stop word removal 🚫 ("the", "is").
2. Stemming/Lemmatization 🌱 ("running" → "run").
3. Remove digits/punctuation 🔢 ("Price: $100" → "price").
4. Lowercasing 🔠 ("Hello" → "hello").

---

7. With examples, explain the differences between segmentation and lemmatization.? 🔄

- **Segmentation**: Splitting text into sentences/words.
  - Example: "Hello world!" → ["Hello", "world", "!"].
- **Lemmatization**: Reducing words to base form.
  - Example: "better" → "good", "meeting" → "meet".

---

8. What is the difference between code mixing and transliteration? 🌐

- **Code Mixing**: Mixing languages in one sentence.
  - Example: "I love the *halwa* (حلوى)!"
- **Transliteration**: Writing non-English words in English letters.
  - Example: "Namaste" (नमस्ते).

---

9. Describe the concept coreference resolution.? 🔗

- **Goal**: Link pronouns to their nouns.
  - Example: "John left. **He** forgot his bag." → "He" = John.

---

10. Explain the feature engineering for classical NLP versus DL-based NLP? 🤖

- **Classical ML** 📊: Handcrafted features (TF-IDF, BoW).
- **Deep Learning** 🧠: Learns features automatically (BERT embeddings).

---

11. How to combine heuristics directly or indirectly with the ML model? 🔄

1. **Heuristics as Features**: Use regex counts in ML models.
2. **Pre-Process Filter**: Block spam domains before ML.

---

12. What is the difference between models ensemble and stacking? 🤝

- **Ensemble**: Combine predictions (e.g., majority vote).
- **Stacking**: Feed Model1's output into Model2.

---

13. Which modeling technique can be used in the following cases: small data, large data, poor data quality, and good data quality? 🛠️

- **Small Data**: Rule-based/ML (Naive Bayes).
- **Large Data**: Deep Learning (LSTM, BERT).
- **Poor Quality**: Clean data first!
- **Good Quality**: Use cloud APIs (e.g., Google NLP).

---

14. What is the difference between intrinsic and extrinsic evaluation? 📈

- **Intrinsic**: Model metrics (accuracy, F1).
- **Extrinsic**: Business impact (e.g., user time saved).

---

15. What are the metrics that can be used in: classification, measuring model quality, information retrieval, predication, machine translation, and summarization tasks.? 🎯

- **Classification** 🗂️: Accuracy, F1 score, Precision.
- **Model Quality** 🔍: Perplexity, Confusion Matrix.
- **IR** 🔍: MRR, MAP.
- **Prediction (Regression)** 🔮: RMSE, MAE.
- **Machine Translation** 🌐: BLEU, METEOR.
- **Summarization** ✂️: ROUGE, BERTScore.

---

16. Describe deploying, monitoring, and updating phases of NLP the pipeline.? 🚀

- **Deploy**: As API (Flask/FastAPI).
- **Monitor**: Track accuracy drops 🚨.
- **Update**: Retrain with new data 🔄.

---

17. Explain how the NLP pipeline is different from a language to another? 🌍

- **High-Resource** (English): Use BERT.
- **Low-Resource** (Swahili): Collect data + transfer learning.
- **CJK** (Chinese): Special tokenizers (Jieba).

---

18. Describe the NLP pipeline for ranking tickets in a ticketing system by Uber.? 🚕

1. **Pre-Process**: Clean text → remove stop words.
2. **Feature Engineering**: TF-IDF + topic modeling.
3. **Modeling**: Rank solutions using text + trip data.
4. **Result**: Saved millions by speeding ticket resolution! 💰

إِنَّ اللَّهَ وَمَلَائِكَتَهُ يُصَلُّونَ عَلَى النَّبِيِّ ۚ يَا أَيُّهَا الَّذِينَ آمَنُوا صَلُّوا عَلَيْهِ وَسَلِّمُوا تَسْلِيمًا (56)