**Creative Industries**

**Task for Distant Contest Stage**

To create an algorithm generating an entire video story (video or music) in return to a text line and genre. Such video story shall be of artistic value and have its video, voice(s), and music track synchronized.

**Preamble**

For a long time, there has been an opinion that algorithms cannot produce items of art. Today, artificial intelligence technologies are gaining popularity in creative industries: people use them to write music and texts and draw pictures.

We suggest you going even further and creating a solution capable of generating a complex and integrated creative work through various algorithms synchronized. For this, you will face the generative (including multimodal) models, understand their specifics, and unite them into a single algorithm. Such solutions might drastically simplify or even completely automate a costly process of creative production and visualization.

**Task for Distant Contest Stage**

You need to produce a solution capable of generating an integrated video story of artistic value. This could be:

- Video with the text voiced and music and video supplemented to illustrate the contents of the story

- Music video – a track (song with lyrics) and a supplementing video

A user request will contain the name and genre of a work to be created:

- Genres to choose for a video story: fable, epic poem, fairy tale, short novel, comedy, tragedy, drama, and horror.

- Genres to choose for a music video: rap, rock, pop, jazz, electro, retro, opera, and musical.

You need to create algorithms for the following tasks to be then synchronized and united into a final solution:

- Text generation

- Text voicing

- Music creation; and

- Video generation.

Texts may be generated in any language. For a multi-language solution, we will assign extra points (see Assessment Criteria below).

**Baseline**

You will be provided a baseline solution from the task setters for the story generation, as well as some alternative technologies in the form of extra notebooks in Google Colab. You may find all the contest materials at GitHub: <https://github.com/sberbank-ai/DigiTeller>

**Solution and Deliveries Format for Distant Contest Stage**

You need to submit your solution for the contest as follows:

* The solution code (preferably the link to a notebook in Google Colab) and trained models (if any) in the form of, for instance, saved checkpoints.
* A link to the generation results (3 to 5 samples).
* The video presentation of the code. The presentation shall be in the form of a video not exceeding 5 minutes. The video shall be in the pitch format presenting the skills obtained and reasoning for the technology solutions used.

**Evaluation criteria for the qualifying round:**

*The final score will consist of the sum of points from the following algorithm models.*

***IMPORTANT:*** *Each model contains underlined points with key criteria. If these criteria are not fulfilled, 0 points are given for the entire model to which the point belongs.*

***1. Model for generating text by name and genre***

* *Consistence of the generated text with the given name and genre – 0-25 points  
  Genre consistence is assessed only in the video story format. The genre of the music video is assessed by evaluating the song generation model.*
* *The length of connected text (as read by your text2speech model or some other model, see Model 2/ Notes):*

*- less than 2 minutes - 0 points*

*- from 2 minutes (inclusive) to 3 minutes – 2 points*

*- from 3 minutes (inclusive) to 4 minutes – 5 points*

*- from 4 minutes (inclusive) to 5 minutes – 7.5 points*

*- over 5 minutes – 10 points*

* *Support for two languages ​​– 10 points, three or more – 15 points*

***Maximum points for this model = 50***

***2. Model for generating voice and music***

*2 variants are possible for this model:*

1. *generation of a background music track and a voice track where the previously generated text is voiced;*
2. *generation of a full-fledged song, where the music and voice tracks must be fully synchronized with each other.*

*Each of the variants has its own rating system and maximum number of points.* ***You can also do both variants and get the sum of their points.***

***Variant 1.***

*2.1.1 Model for generating background music*

* *Quality of the generated music – 0-10 points*
* *Its ability to convey the mood of the story at different moments – 0-10 points*

*2.1.2 Model text2speech*

* *The quality of the generated voice, consistence with the text – 0-10 points*
* *Multi-voice generation (generation of voices for each character, if there are several of them in the story) – 0-5 points*
* *Support for two languages – 10 points, three or more – 15 points*

***Maximum points for this variant = 50***

***Variant 2.***

*2.2.1 Song generation*

* *Quality of the voice track and its Consistence with the text - 0-10 points*
* *Quality of the music track – 0-10 points*
* *Synchronization with the voice track - 0-35 points*
* *Generation of multi-voice in case there are dialogues in the song – 0-10 points*
* *Consistence of music and voice tracks with the set genre 0-10 points*
* *Support for two languages – 10 points, three or more – 15 points*

***Maximum points for this variant = 90***

***3. Model for generating video***

*This model also has two different variants:*

1. *generation of pictures by lines of text and their composition (e.g. by interpolation or morphing)*
2. *generation of a full video.*

*Each of the variants has its own rating system and the maximum number of points.* ***You can also do both variants and get the sum of their points.***

*3.1 – Generation of pictures by lines of text and their composition:*

* *Consistence of each frame and the meaning of the text – 0-20 points*
* *Consistence of the frame style with the storytelling style – 0-10 points*
* *Composition quality of the generated frames with each other into a single video – 0-10 points*
* *Resolution of the generated clip:*

*- 512х512 – 0 points*

*- 1024х1024 – 2.5 points*

*- 2048x2048 – 5 points*

*- 4096x4096 – 7.5 points*

*- 8192x8192 – 10 points*

***Maximum points for this variant = 50***

*3.2 – Generation of a full video\*:*

* *Consistence between visuals of each frame, the meaning of the text, genre and style of the music and sound tracks– 0-110 points*
* *Resolution of the generated clip:*

*- 512х512 – 0 points*

*- 1024х1024 – 2.5 points*

*- 2048x2048 – 5 points*

*- 4096x4096 – 7.5 points*

*- 8192x8192 – 10 points*

*\* The format of video clips can be very diverse. For example, you can generate a person (or a virtual character) who will “sing” this song and move to it (on a virtual stage). To do this, you can use the following ideas:*

*a) “sing” the song yourself (that is, shoot a video with your participation). Then use FaceSwap (BodySwap), and in the resulting video, replace yourself with this character, but preserving movement (see the overview of* [*FaceSwap*](https://www.notion.so/FaceSwap-350382dc415d4dff9fe282500a524e13) *methods).*

*б) generate (or set manually) the position of the character in space (so-called "skeleton"). And then generate a videorealistic version of this character (*[*example*](https://people.mpi-inf.mpg.de/~mhaberma/projects/2021-ddc/)*).*

*You can use a slideshow of generated pictures or abstract images in the background.*

*Any original ideas are accepted (the idea above is given only as an example).*

***Maximum points for this variant = 120***

*The final score will be given by an expert jury.* ***The maximum possible score is 360 points****.*

***Bonus:****Publish your results on social media with hashtags #SberDigiTeller, #SberAI, #AIIJC, with more than 100 likes, 10 points are additionally awarded, more than 1000 likes - 20 points.*

*Notes:*

1. ***The duration of the final video should not be less than 1 minute. Solutions less than 1 minute will not be considered (default to 0 points).***
2. ***The maximum possible score is 380, when taking the social media bonuses in account.***